Cet ouvrage est diffusé en accès ouvert dans le cadre du projet OpenEdition Books Select.
Ce programme de financement participatif, coordonné par OpenEdition en partenariat avec Knowledge Unlatched et le consortium Couperin, permet aux bibliothèques de contribuer à la libération de contenus provenant d'éditeurs majeurs dans le domaine des sciences humaines et sociales.
La liste des bibliothèques ayant contribué financièrement à la libération de cet ouvrage se trouve ici :
https://www.openedition.org/22515.

This book is published open access as part of the OpenEdition Books Select project.
This crowdfunding program is coordinated by OpenEdition in partnership with Knowledge Unlatched and the French library consortium Couperin. Thanks to the initiative, libraries can contribute to unlatch content from key publishers in the Humanities and Social Sciences.
Discover all the libraries that helped to make this book available open access: https://www.openedition.org/22515?lang=en.

QU'EST-CE QU'UNE **ARCHIVE DU WEB ?**

FRANCESCA MUSIANI,
CAMILLE PALOQUE-BERGÈS,
VALÉRIE SCHAFER,
BENJAMIN G. THIERRY

QU'EST-CE QU'UNE **ARCHIVE DU WEB ?**

OpenEdition press

En couverture : *Empty honey background*
© jonnysek
Conception graphique : Veronica Holguín - Collectif Surletoit

Correction : Solenne Louis
Suivi et coordination éditoriale : Cédric Gaultier et Caroline Terrier
Cet ouvrage a été mis en page grâce à Métopes. Méthodes et outils pour l'édition structurée

Francesca Musiani, Camille Paloque-Bergès, Valérie Schafer, Benjamin G. Thierry : *Qu'est-ce qu'une archive du Web ?*
Collection « Encyclopédie numérique », 2019
[En ligne] http://books.openedition.org/oep/8713

ISBN papier : 979-10-365-0368-9
ISBN électronique : 979-10-365-0367-2

REMERCIEMENTS

Cet ouvrage doit beaucoup à OpenEdition Press. Nous remercions les directeurs de la collection « Encyclopédie numérique », Marin Dacos et Christian Jacob, de leur confiance et Cédric Gaultier et Caroline Terrier de leur aide au cours du projet.

Parce que cette collection permet un format court, pédagogique, ouvert – en termes d'accès, mais aussi de public –, en visant une audience large tout en accueillant l'expertise, et parce qu'elle offre la possibilité d'intégrer les potentialités du Web et notamment de l'hypertextualité, le projet nous a séduits.

Il avait aussi séduit Louise Merzeau. Faire ce livre sans elle reste notre regret, mais le faire pour elle nous tenait à cœur.

Ces remerciements ne seraient pas complets si nous n'évoquions pas le soutien de l'ANR au projet Web90 de 2014 à début 2018, les collègues qui y ont participé avec nous (Mélanie Dulong de Rosnay, Fanny Georges, Hervé Le Crosnier et Stéphanie Le Gallic), et l'ISCC qui l'a hébergé, ainsi que les échanges féconds noués au fil des années avec la BnF et l'Ina, en particulier avec les équipes du DL Web, et avec le groupe de recherche européen dédié aux archives du Web, RESAW, fondé par Niels Brügger.

INTRODUCTION

Pour les vingt ans de la fondation Internet Archive, créée en 1996 et pionnière dans l'archivage du Web aux États-Unis, ainsi que pour les dix ans du dépôt légal du Web en France, la Bibliothèque nationale de France (BnF) et l'Institut national de l'audiovisuel (Ina) coorganisaient en 2016 un colloque intitulé « Il était une fois dans le Web : 20 ans d'archives de l'internet en France ». S'il fallait conter l'épopée de l'archivage du Web, sans doute y croiserait-on quelques preux protagonistes partis en quête d'un archivage du Web mondial ou exhaustif, des défis à affronter – humains, législatifs ou techniques – où cohabiteraient notamment captcha, robots.txt et droits d'auteurs, des issues heureuses aussi, tel le mariage prometteur de l'archive du Web et de la recherche. Roman chevaleresque, conte de fées, livre dont vous êtes le héros ? Si sans doute l'archive du Web pourrait s'y prêter, ce livre se contentera d'une mise en intrigue plus classique et d'initier le lecteur aux enjeux des archives et de l'archivage de la Toile. Il raconte l'archive de manière pragmatique, en essayant de rendre palpable à la fois sa fabrique (technique, institutionnelle, juridique), les évolutions qu'a connues cet archivage en une vingtaine d'années et la relation que les acteurs de l'archivage du Web entretiennent avec les publics. Ce tour d'horizon ne serait pas complet sans s'intéresser à la manière dont ces archives, outre leurs qualités patrimoniales, peuvent aujourd'hui être exploitées par le monde de la recherche. Il sera donc question d'archivage, d'outils et de métadonnées, mais aussi d'héritage culturel, de géopolitique ou encore d'éthique, tant les enjeux qui touchent au patrimoine et à l'archive ne peuvent être dissociés d'enjeux de mémoire et d'histoire aux multiples parties prenantes.

Organisé en quatre parties, notre propos suit l'archive du Web depuis sa naissance et sa conservation, dans les deux premières parties, jusqu'à son exploitation.

Lorsque l'on évoque l'archive du Web, il faut se figurer un objet singulier, interactif, fluide et non figé. Mais aussi une archive qui, bien qu'elle ressemble au Web du passé, n'en est pas la copie conforme et peut selon les fonds prendre des formes distinctes, enchâssées dans des interfaces, supportées par des techniques qui livrent des résultats visuellement différents. L'exemple le plus évident, sur lequel nous reviendrons, est celui de l'archivage du réseau social numérique Twitter : ici les différences d'archivage entre les deux institutions françaises en charge du dépôt légal du Web en France, la BnF et l'Ina, sont visibles à l'œil nu. À la BnF, les archives de Twitter s'apparentent à des captures d'écran, tandis que l'Ina a fait le choix d'une collecte fondée sur des données davantage brutes, sans capturer les images de fond. Mais au-delà, toute archive du Web véhicule des choix, des arbitrages. Ces choix ne sont pas seulement techniques, mais aussi profondément humains et sociétaux, voire politiques, ce qui est sans aucun doute le lot traditionnel d'autres types d'archives.

Avec le patrimoine pléthorique du Web en cours de constitution, outre la question de la masse (plus de 345 milliards de pages web archivées depuis 1996 par la fondation Internet Archive[1]) se pose la question de la collecte, largement automatisée. Car les archives du Web introduisent bien des ruptures, que ce soit dans la notion même d'archive, ou dans les pratiques des archivistes et des chercheurs, même si on peut également y voir des continuités, comme nous le montrerons.

Automatisée, la collecte des archives du Web l'est à partir de périmètres négociés, et donc de choix humains. De cette curation, au moins initiale, dépend la représentativité de l'archivage, par elle se lit aussi l'inégale valeur accordée aux matériaux nativement numériques, archivés ou exclus de la collecte. Outillée au besoin, l'analyse des archives du Web l'est au service de questions posées par le chercheur, et là encore de choix humains. Ce sont aussi ces agentivités et interactions humaines et techniques que cet ouvrage propose de découvrir.

Ce projet est né d'une double volonté collective : celle de prolonger une initiative pensée avec Louise Merzeau quelques mois

1. Voir https://archive.org/web/.

avant sa disparition, elle qui, dans les ateliers du DL Web Ina[2] avec Claude Mussou ou encore dans notre projet ANR Web90[3] et au fil de ses écrits, a tant fait pour penser, mais aussi pour faire connaître l'archive du Web. Le désir également de partager notre « goût » de l'archive du Web, alors que nous la prenons depuis plusieurs années comme source et objet d'étude. Nous espérons que cet ouvrage convaincra d'ailleurs le lecteur de considérer de manière indissociable la création de l'archive et son analyse et qu'il lui fournira des clés pour cela. Et peut-être sera-t-il même tenté de reconsidérer ensuite avec un regard décalé d'autres archives que celles nativement numériques ?

2. Voir le blog des ateliers : http://atelier-dlweb.fr/blog/.
3. https://web90.hypotheses.org.

DES ARCHIVES COMME LES AUTRES ?

Les discussions sur les archives du Web, en particulier quand elles ont lieu entre historiens, débouchent régulièrement sur la question de la rupture ou continuité de ces archives avec les précédentes. Et bien sûr la réponse n'est pas univoque. Certains éléments peuvent être rapprochés de situations antérieures : les enjeux liés à l'exhaustivité et la représentativité des fonds ne sont pas nouveaux, comme ceux sur l'authenticité des documents ou sur l'outillage numérique de l'analyse (par exemple pour l'exploitation de séries statistiques ou de sources audiovisuelles). La masse et la surabondance documentaires sont connues de beaucoup d'historiens du contemporain, de même que les « éphémères » pour ceux qui s'intéressent aux cultures vernaculaires ou aux tracts politiques. Toutefois, des différences certaines existent, qui peuvent même inviter à remettre en question la pertinence de l'emploi de la notion d'archive. Si institutions et internautes parlent d'archives du Web, Bruno Bachimont (2017b) revenant sur l'organisation des traces dans le cadre de l'archive, de la bibliothèque et du centre de documentation y voit plutôt des collections. Il rappelle que l'archive, elle, est conçue pour constituer « une preuve sur ce qui s'est passé » (*ibid.*) : « l'enjeu est de pouvoir conserver les documents permettant de renseigner, reconstituer et prouver les activités de l'institution concernée, les événements auxquels elle a pris part. Aussi l'enjeu est-il de garder le plus possible le lien organique entre le document et l'activité qui l'a produit, pour que l'examen de l'effet qu'est l'archive permette de remonter à la cause qu'est l'événement » (*ibid.*). À l'inverse, « 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 (*ibid.*). Inscrites

dans le monde des bibliothèques et dans le cadre d'un dépôt légal qui conserve des œuvres de l'esprit davantage que des traces d'activité, les archives du Web s'apparentent ainsi plus à des collections. L'archive du Web invite donc à (re)penser le rapport du chercheur comme des professionnels de l'archivage et des bibliothèques aux données, aux documents, aux collections et aux archives.

Aussi c'est en termes de patrimoine, de statut de ces fonds, mais également de contexte que les archives du Web sont présentées dans cette première partie, qui ne pouvait manquer bien sûr de s'ouvrir par leur courte mais déjà riche histoire.

Une brève histoire de l'archivage du Web

On est bien entendu tenté de faire commencer l'histoire des archives du Web en 1996, avec la création de la fondation Internet Archive par Brewster Kahle[1]. Sans remonter en France à la création du dépôt légal sous François I^er^ (1537), ou reprendre dans le détail une chronologie qui a vu après les imprimés son extension aux matériaux numériques tels les vidéogrammes et documents multimédias composites (1975), puis aux multimédias, logiciels et bases de données (1992) (Oury in Cohen et Verlaine, 2013), on pourrait aussi faire débuter cette histoire en 1989. Pas seulement parce que c'est le moment où le Britannique Tim Berners-Lee commence à travailler au projet de ce qui deviendra le World Wide Web, qui connaîtra dans la décennie 1990 une popularisation sans précédent, mais aussi parce qu'en 1989 Brewster Kahle invente un système de publication sur Internet, le WAIS (Wide Area Information Server) et fonde WAIS Inc., qu'il revend à America Online (AOL) en 1995. L'année suivante, lancé dans la voie des technologies internet et web et fort de ce succès, il fonde Internet Archive et Alexa, entreprise qu'il vend à Amazon.com en 1999 :

1. Au-delà de la portée symbolique du dixième anniversaire du DL Web et des 20 ans d'Internet Archive, célébrés de concert par la BnF et l'Ina en 2016, le colloque « Il était une fois dans le Web », organisé à cette occasion, offrait un regard rétrospectif mais aussi prospectif sur l'archivage du Web (dont certains éléments et témoignages peuvent être retrouvés dans le carnet de recherche Web Corpora de la BnF : https://webcorpora.hypotheses.org/200#more-200).

> « Ce qui n'était au départ qu'un simple projet de recherche va vite devenir une société basée à San Francisco, à l'origine dès juillet 1997 d'un outil commercial appelé Alexa. Cet outil permet de "butiner", rapatrier et indexer un nombre important de pages et de donner des indications sur leur fréquentation, le renouvellement, le nombre de liens, mais surtout il permet de donner accès aux versions précédentes des sites archivés par Internet Archive. » (Chaimbault, 2008)

Et même si l'on fait commencer cette histoire en 1996, cette année ne se limite pas à la fondation d'Internet Archive. Trois autres initiatives émergent : une en Australie, une archive tasmanienne – également issue d'une initiative australienne –, et enfin Kulturarw3 en Suède. Seules cinq autres initiatives d'archivage du Web naîtront dans les six années suivantes, avant que 2003 ne marque un décollage (Gomes *et al.*, 2011). Mais déjà toutes ces initiatives donnent le ton de la diversité de l'archivage du Web : à « l'approche intégrale » d'Internet Archive qui se donne pour ambition d'archiver le Web mondial à ses débuts, répond une « approche exhaustive » de la part de la Bibliothèque royale de Suède, qui cherche à conserver tout le .se[2], tandis que l'Australie opte pour une « approche sélective ». Des « approches thématiques » ou encore « combinées » viendront dans les années suivantes compléter cette typologie (Chaimbault, 2008), ce qui montre bien à quel point le périmètre d'archivage peut varier. Quant à la France, dès la fin des années 1990 elle s'intéresse à la question, sans toutefois entrer encore officiellement sur la scène de l'archivage du Web.

Le début des années 2000 est marqué par deux étapes majeures, en termes de conservation comme d'accessibilité. En 2001 naît la Wayback Machine[3] d'Internet Archive, porte d'accès en ligne aux archives de la fondation. Et en 2003 une charte de l'Unesco sur la conservation du patrimoine

2. Nom de domaine national de premier niveau de la Suède.
3. Pour accéder à la Wayback Machine : https://archive.org/web/.

numérique[4] fait explicitement allusion au patrimoine nativement numérique. Mentionnant à deux reprises le *born-digital heritage* ou patrimoine « d'origine numérique » (article 1er et article 7), la charte le distingue du patrimoine numérisé en ce qu'il existe sous forme numérique dès son origine (c'est le cas des sites web, des bases de données, etc.), alors que le second a subi un processus de numérisation. Si la reconnaissance du patrimoine numérique – et notamment du patrimoine d'origine numérique – est à mettre en relation avec le développement important au cours du XXIe siècle des communications en réseau, elle doit aussi être mise en lien avec des tendances qui depuis une vingtaine d'années ont pu faire parler de véritable « explosion patrimoniale » (Nora, 1996), diversifiant les objets reconnus comme faisant partie du patrimoine (notons, en 2003 également, la reconnaissance du patrimoine culturel immatériel, voir Severo et Cachat, 2017). La place croissante de la culture et de la mémoire techniques (Bouvier, Polino et Varaschin, 2010) ou encore la progressive patrimonialisation de la communication (Paloque-Bergès et Schafer, 2015) ont aussi joué un rôle dans ce mouvement.

L'année suivante, en 2004, est créé l'International Internet Preservation Consortium[5]. L'IIPC rassemble au départ 12 membres, une cinquantaine aujourd'hui, soit une bonne partie des institutions qui se sont investies ces dernières années dans l'archivage du Web (voir la liste des initiatives d'archivage du Web rassemblées sur Wikipedia[6]). Les missions de l'IIPC sont dès l'origine de favoriser la collaboration internationale, mais des priorités peuvent ensuite être distinguées au fil de ses presque quinze années d'existence. Aux réflexions sur la compatibilité des formats et une politique de normalisation fondée sur le format WARC à la fin des années 2000 ou l'adoption du modèle OAIS (Open Archival Information System) dédié à l'archive numérique, s'ajoutent depuis quelques années des réflexions sur le traitement des données sauvegardées et la manière d'assurer leur intégration dans les collections des bibliothèques (Gebeil, 2014). Car de la Bibliothèque royale du Danemark à la Bibliothèque

4. Voir : http://portal.unesco.org/fr/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html.
5. Voir le site de l'IIPC : http://netpreserve.org.
6. https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

du Congrès aux États-Unis (Library of Congress, ou LoC), en passant par la British Library ou encore la BnF, de nombreuses bibliothèques se sont investies dans l'archivage du Web.

En France le dépôt légal, à savoir « l'obligation pour tout éditeur, imprimeur, producteur, importateur de déposer chaque document qu'il édite, imprime, produit ou importe en France à la BnF ou auprès de l'organisme habilité à recevoir le dépôt en fonction de la nature du document[7] » est élargi aux publications sur Internet (sites institutionnels ou personnels, revues d'accès gratuit ou payant, blogs, sites commerciaux, plateformes de vidéos, etc.) depuis la loi du 1er août 2006 relative au droit d'auteur et aux droits voisins dans la société de l'information (DADVSI). Toutefois, contrairement au dépôt traditionnel, l'éditeur de contenu n'a pas à accomplir de démarche active de dépôt. En effet, ce sont la BnF et l'Ina qui se sont vu confier l'archivage du Web, dans le cadre de leurs périmètres respectifs. L'Ina conserve des contenus qui relèvent de l'audiovisuel, tandis que la BnF prend en charge « le reste » d'un ensemble qui ne se limite pas au .fr, mais intègre des extensions territoriales (par exemple le .re) et les contenus produits par des Français ou des auteurs domiciliés en France, dont les adresses sont en .com, .org, etc. Près de 4,5 millions de sites sont ainsi collectés par la BnF chaque année. D'autres pays ont adopté des mesures proches, faisant entrer les publications en ligne dans le cadre du dépôt légal (en 2013 pour le Royaume-Uni, en 2017 pour toutes les publications numériques en Belgique).

En outre, dans la décennie 2010 les réseaux socionumériques (RSN) commencent à susciter l'intérêt et la Library of Congress passe un accord avec Twitter pour conserver les archives des tweets. L'Ina se met à collecter Twitter à partir de 2014, toujours dans le cadre de son périmètre puisqu'il s'agit de suivre des comptes liés à l'audiovisuel et aux professionnels du secteur français. L'année précédente, l'institut avait commencé la captation des radios web et dès 2010 celle des plateformes vidéos comme YouTube ou Dailymotion : dans un souci de cohérence et de continuité des collections, l'Ina cherche à suivre de près les mutations des pratiques de diffusion mais

7. Voir http://www.bnf.fr/en/professionnels/depot_legal.html.

aussi de réception de l'audiovisuel. En effet, le développement de plateformes en ligne et celui de la participation aux réseaux socionumériques invitent à penser ces pratiques du « deuxième écran » et à suivre des contenus qui participent pleinement du périmètre audiovisuel.

Chaque institution a ainsi des contraintes, enjeux, motivations spécifiques, mais aussi ses rythmes propres. La BnF distingue plusieurs étapes dans l'histoire de son archivage[8] : la période 1999-2004 ou le temps des expérimentations ; 2004-2007 ou la mise en place d'un « modèle intégré[9] », stabilisé juridiquement par la loi DADVSI ; et 2007-2012 avec la réalisation d'un cycle d'archivage complet. À ces trois périodes, on peut en ajouter une plus récente : dans le cadre de son projet WebCorpus, inscrit au plan quadriennal de recherche 2016-2019, la BnF pense à élaborer un service de fourniture de corpus aux chercheurs (Moiraghi, 2018), mobilisant notamment des technologies de fouille de textes et de données, ainsi que de nouvelles possibilités d'exploitation des fichiers issus de la capture et de l'indexation automatiques des sites web.

Le cas européen

En France, l'État a, en créant le dépôt légal du Web, consacré la place d'un « tiers neutre qui garde la mémoire de ce qui est publié sur le Web sans en faire un objet commercial » (Oury in Cohen et Verlaine, 2013). Mais qu'en est-il des autres pays européens, et des institutions européennes elles-mêmes ?

Arquivo.pt, qui cherche à conserver le Web portugais et les informations publiques en ligne relatives à la communauté portugaise, compte actuellement plus de 100 000 utilisateurs, dont la moitié hors Portugal. Née en 2008, cette infrastructure est accessible en ligne, contrairement à d'autres fonds auxquels on ne peut accéder que depuis des bibliothèques ou sites dédiés

8. Voir sur le site de la BnF : http://www.bnf.fr/fr/professionnels/archivage_web_bnf/a.depot_legal_internet_histoire.html. Voir également Aubry, 2010.
9. « Il s'agissait de réaliser conjointement des collectes larges, "aveugles", du domaine français, conjuguées avec des collectées, plus profondes ou plus fréquentes, de sites sélectionnés par des bibliothécaires » : http://www.bnf.fr/fr/professionnels/archivage_web_bnf/a.depot_legal_internet_histoire.html.

après avoir reçu une accréditation recherche. L'initiative a également une ambition de recherche, avec le développement d'outils et la publication de plusieurs dizaines d'articles en accès libre. La dimension de développement d'outils intégrés est également présente en Suède, où le programme Kulturarw3, qui existe depuis 1996, dispose de son propre système de stockage et d'accès.

Le projet d'archivage du Web en Belgique est porté par une initiative de recherche – chapeautée par la Bibliothèque royale et les Archives nationales, avec la participation de plusieurs universités – et il est tout récent : le projet PROMISE[10] voit en effet le jour en 2017 et œuvre actuellement à un pilote pour archiver le Web belge, sur la base d'une étude des bonnes pratiques dans d'autres pays.

En miroir du cas français, l'archivage du Web aux Pays-Bas est assuré par deux institutions : la Koninklijke Bibliotheek, qui a une mission d'identification et de sauvegarde sélectives de sites néerlandais ayant une valeur culturelle et scientifique ; et l'Institut néerlandais du son et de la vision qui a débuté son investissement dans l'archivage en 2008 pour le périmètre audiovisuel.

Selon les pays européens, l'amplitude et les critères de la collecte des sites varient. Le cas espagnol est intéressant : les archives web de ce pays sont entretenues par la bibliothèque nationale avec la collaboration d'un réseau de bibliothèques régionales (une approche également adoptée par la Suisse) et sont le résultat d'un mélange de collectes inclusives et sélectives.

D'autres pays adoptent également ce critère mixte : par exemple, en Finlande, la bibliothèque nationale conduit une collecte annuelle de tous les domaines .fi et des serveurs web qui se trouvent sur le territoire finlandais, mais au-delà de ces collectes, elle sélectionne manuellement des sites web qui lui semblent particulièrement pertinents (sites d'information, culturels, etc.). C'est également le cas du Luxembourg, qui conduit deux fois par an des collectes amples ainsi que des collectes plus sélectives, notamment à l'occasion d'événements particuliers, par exemple des élections. L'approche est la même en

10. https://promise.hypotheses.org/.

Croatie, qui a commencé en 2004 avec une collecte sélective, ensuite élargie à des collectes annuelles complètes du domaine .hr et des collectes thématiques ou/et liées à des événements « d'intérêt national ». Au contraire, en Irlande, la bibliothèque nationale opte pour une approche uniquement sélective de sites « d'importance scientifique, culturelle et politique ».

Un autre aspect qui varie selon les pays est la modalité d'accès aux archives du Web. Au Royaume-Uni, l'archivage du Web est du ressort à la fois de la British Library, dont une partie des collections est accessible en ligne (UK Web Archives) et des archives parlementaires, également en ligne[11]. Si Arquivo.pt, cité précédemment, propose également ses ressources en ligne et en accès libre, comme l'Islande ou la Croatie, d'autres, pour des raisons notamment de droit d'auteur, proposent comme la BnF de limiter l'accès aux archives du Web à partir des lieux physiques de l'institution. C'est le cas de l'Allemagne, qui, au-delà d'une archive web réunie et hébergée par le Bundestag, dispose d'une archive qui résulte d'une collecte sélective, conduite par l'entreprise oia GmbH, dont l'accès est restreint aux salles de lecture de la Bibliothèque nationale allemande. Dans certains cas, les modalités d'accès ont évolué : en Estonie, une première loi sur le dépôt légal de 2006 a permis à la bibliothèque nationale de récupérer régulièrement une sélection de sites web nationaux, que cette dernière a d'abord rendus disponibles en libre accès ; cependant, une nouvelle loi de 2017 a rendu l'accès possible seulement avec la permission des ayants droit. Une loi sur le dépôt légal régit également les collectes espagnoles, rendant les sélections de sites web disponibles pour le public « en observant les règles du droit d'auteur ».

Au-delà de ces archivages nationaux, conscient de la valeur de ce patrimoine nativement numérique prompt à disparaître ou changer, l'Office des publications européennes a débuté, en 2013, un archivage tourné vers les sites web d'agences et d'institutions européennes[12], dont la plupart sont hébergées par le domaine europa.eu.

11. http://webarchive.parliament.uk.
12. https://www.eui.eu/Research/HistoricalArchivesOfEU/WebsitesArchivesofEUInstitutions.

Une composante du patrimoine nativement numérique

On peut parler de « patrimoine d'origine numérique » ou de « patrimoine nativement numérique », plus proche de l'expression « *born-digital heritage* ». Plus restreint que le patrimoine numérisé, qui s'étend aux ressources analogiques converties sous forme numérique, il embrasse les matériaux et formats produits initialement sous forme numérique, incluant « les textes, les bases de données, les images fixes et animées, l'audio, le graphisme, le logiciel et les pages web ». L'idée d'un « patrimoine d'origine numérique » comme nouveau legs de la mémoire mondiale est officiellement reconnue et stimulée par la charte de l'Unesco de 2003 sur la conservation du patrimoine numérique, qui s'inscrit dans la continuité du programme « Memory of the World » initié par l'Unesco en 1992. Cet acte de naissance du patrimoine nativement numérique est accompagné d'une double injonction. Tout d'abord, un appel à la coopération entre les différents corps professionnels, publics ou privés, spécialisés dans le numérique (développeurs de logiciel, créateurs, éditeurs, producteurs et distributeurs) et les institutions de préservation patrimoniale (bibliothèques, archives, musées, etc.). Ensuite, la reconnaissance de la priorité à donner à cet aspect spécifique, natif, du patrimoine nativement numérique, tout aussi bien en raison du caractère inédit de sa préservation que de l'urgence de sa collecte.

Au-delà d'une liste de ressources types, que recouvre la réalité du patrimoine nativement numérique ? Il prend forme à la fois dans la préservation des technologies d'information, des objets numériques créés lors de leur utilisation, ainsi que de l'information que ces objets transportent, comme le définit Ken Thibodeau (Unesco, 2012). Les archives du Web sont en cela tributaires des limites sinon floues, du moins fluctuantes, entre ces trois dimensions. Le numéro que *La Gazette des archives* a consacré à « Archives et Internet » en 2007 (Verry, 2007) en témoigne : il présente des travaux aussi bien sur les sites web des institutions d'archivage (à la fois vecteurs d'information et interfaces de communication avec les publics), que sur la conception des outils, les usages ou le design d'expérience.

Le patrimoine de l'informatique a pavé la voie et contribue fondamentalement à la « fabrique du patrimoine numérique » (Musiani et Schafer, 2017), aussi bien au niveau matériel qu'immatériel. Les premières initiatives patrimoniales viennent de l'intérieur du domaine. En effet, elles sont déployées par les acteurs de terrain, premiers concernés par la préservation d'une mémoire professionnelle et/ou ludique des machines numériques. Aux associations d'anciens professionnels de l'informatique comme ACONIT (Association pour un conservatoire de l'informatique et de la télématique) ou la FEB (Fédération des équipes Bull) en France, se sont ajoutées des initiatives institutionnelles s'inscrivant dans une tradition muséale, avec des collections spéciales, comme au Musée des arts et métiers français, ou des établissements dédiés, comme le Computer History Museum aux États-Unis. En France, c'est l'Institut national de recherche en informatique (Inria) qui porte le grand projet d'une archive mondiale du logiciel, Software Heritage, destinée à préserver les codes sources. Des organisations clés dans le domaine de l'internet et du Web comme l'Internet Engineering Task Force (IETF) ou le World Wide Web Consortium (W3C) déploient très tôt une politique de valorisation et d'accessibilité aux archives nativement numériques pour documenter leur propre histoire, de leur contribution scientifique et technique à Internet à leur participation à sa gouvernance – en particulier les forums électroniques qui ont permis de structurer leur travail collectif depuis plus de trente ans. En élargissant quelque peu la perspective, on doit aussi considérer les apports primordiaux des groupes et communautés d'amateurs d'informatique. Les collections d'Internet Archive leur font d'ailleurs une large place, incluant nombre de matériels et logiciels ayant marqué les premières générations d'utilisateurs dès les années 1980 – avec une forte présence, par exemple, de l'univers vidéo-ludique. L'Archive Team, organisation de bénévoles formée en 2009, se spécialise, elle, dans la sauvegarde d'urgence de certains espaces de sociabilité en ligne ayant jalonné l'histoire culturelle du Web et aujourd'hui disparus, comme Geocities ou Friendster.

La reconnaissance d'un patrimoine nativement numérique ne se limite pas aux intérêts de ces publics pionniers, malgré

leur rôle indéniablement moteur. Le patrimoine nativement numérique suscite en particulier l'intérêt réflexif des professionnels du document pour l'évolution de leurs objets, matériaux et outils de travail. Cela peut expliquer la précoce inscription de la sauvegarde des archives nativement numériques dans les services de bibliothèques. Le sens de l'archive nativement numérique se pense d'abord fondamentalement, comme le souligne le chercheur Fabrice Papy, « entre bibliothèque et informatique » (Papy, 2015, p. 32). Les matériaux de l'archivage engagent les professionnels dans une réévaluation de leurs outils de travail et l'expérimentation de nouveaux dispositifs. Par exemple, les techniques de l'interopérabilité viennent répondre, à l'ère des réseaux hypertextuels, aux besoins traditionnels du monde documentaire en matière de standards pour mettre en forme, identifier, et communiquer des documents. L'analyse et le codage (par les langages et formats numériques) de données informatiques et en réseau répondent aux logiques de visibilité et d'accessibilité des contenus sur le Web, en permettant une nouvelle approche des métadonnées documentaires. Le développement de formations pour les documentalistes du XXI^e^ siècle atteste ces nouvelles compétences d'analystes et de programmation, alliées aux sciences de l'information (Niu, 2012). Les archives du Web ne peuvent être envisagées sans la mise en place de dispositifs expérimentaux en matière de logiciels et langages numériques. Ces derniers peuvent s'inspirer de travaux d'équipes de développeurs du Web, en adoptant, ou tout du moins en adaptant les langages et standards issus des entrepreneurs de l'informatique. Par exemple, le projet « 404 no more », collaboration entre Mozilla/Firefox et la fondation Internet Archive, redirige automatiquement vers les collections de cette dernière pour les pages disparues auxquelles on tente d'accéder par le navigateur Firefox. Des technologies similaires ont pu être utilisées dans le projet Memento[13] développé à la Los Alamos National Laboratory Research Library. Dans la même perspective, les logiciels

13. Memento est une extension logicielle que l'on peut greffer à son navigateur, et qui permet de fouiller dans les différentes archives du Web qui acceptent d'afficher leurs données selon un protocole spécifique à Memento. Le but est de pouvoir afficher des anciens contenus comme s'ils étaient encore actifs. http://mementoweb.org/about/.

développés par la fondation privée Internet Archive sont très utilisés par les institutions patrimoniales dans l'archivage du Web, à commencer par le robot d'indexation[14] Heritrix, conçu dès 2003 pour l'archivage du Web en dialogue avec l'IIPC.

Il faut noter deux tournants majeurs dans la conception du patrimoine. D'une part, la progressive valorisation de l'information qu'il peut contenir : ce n'est plus seulement l'enjeu de la mémoire, mais cette dimension d'information qui est mise en avant (Unesco, 2012). D'autre part, la préservation, aux côtés des artefacts matériels, des artefacts immatériels ; et, aux côtés des monuments, de patrimoines de plus en plus diversifiés, notamment un patrimoine lié à la communication (Paloque-Bergès et Schafer, 2015). L'explosion des contenus et outils numériques crée tout autant l'espoir que l'anxiété. Cela découle du constat d'une numérisation exponentielle des activités humaines dans les pays industrialisés, et donc de celui d'une partie de plus en plus grande de l'héritage mondial, comme le relève Wendy Hanamura de la fondation Internet Archive[15]. Ce constat s'accompagne d'un sentiment d'urgence, largement légitimé par le programme « World Memory Heritage » de l'Unesco qui abrite des projets tels qu'« Archives at risk[16] » et qui, en 2012 déjà, rappelait le risque de perte d'autant plus grand que « le numérique est devenu le canal principal de la production et de la transmission de savoir » (Unesco, 2012). Cette anxiété est relayée par les professionnels de l'archive non seulement au niveau des pratiques, mais aussi des droits relatifs à la conservation des documents de mémoire. La mobilisation de l'association des archivistes français en 2013 contre des projets de lois européennes pour formaliser un droit à l'oubli numérique (mobilisation #EUdataP) fournit un exemple intéressant de débat public autour de ce problème.

14. Logiciel qui explore automatiquement le Web, afin de collecter des ressources et ensuite permettre à un moteur de recherche de les indexer. L'aspect « exploration » est souvent appelé *crawling*, d'où le terme également de *robot crawler*.
15. https://venturebeat.com/2015/10/22/the-internet-archive-is-rebuilding-the-wayback-machine-to-make-the-webs-history-easier-to-search/.
16. Une initiative mondiale qui vise à sauvegarder les archives audiovisuelles menacées, en sensibilisant l'opinion, en encourageant les projets de coopération et en s'appuyant sur l'expertise et le soutien des principales organisations représentant les archives audiovisuelles : http://archivesatrisk.com/about/.

En cela, la réflexion sur le patrimoine nativement numérique, et en son sein la question des archives, prépare le terrain à une future archéologie du savoir, qui étudierait les conditions de production des discours et du savoir au sein de dispositifs techniques et sociaux, comme y invitent les chercheurs en archéologie des médias[17] (Parikka, 2013).

Les archives du Web entre rupture et continuité

Il est évidemment tentant de penser les archives du Web avant tout en termes de rupture par rapport à des archives plus « traditionnelles », que ce soit en raison de la masse de données accumulée ou encore de la difficile sélection : la collecte est automatisée, déléguée à des robots, bien qu'ils soient évidemment programmés par des acteurs humains. En archivant un mot-dièse (*hashtag*) de Twitter, comme en programmant un robot pour les collectes hebdomadaires d'un site web de presse par exemple, rien ne garantit le contenu exact qui sera collecté. Bien sûr le périmètre s'appuie sur un cadre législatif pour les dépôts légaux et les choix sont discutés au sein des institutions qui décident de la profondeur ou encore de la récurrence de la collecte d'un site. Mais le périmètre de la collecte est fixé a priori sans savoir exactement quel sera le contenu disponible au moment du passage du robot, ni la valeur des informations recueillies pour le présent et le futur.

Notons d'ailleurs que cette collecte rompt aussi avec la tradition du dépôt légal, ce que relevait Clément Oury à propos de « cette partie du dépôt légal qui, contrairement à celui des imprimés, ne reçoit pas de communication de la part des éditeurs de contenu, mais élabore une cible documentaire, va à sa recherche suivant deux modes principaux de collecte : la collecte large, et les collectes ciblées » (in Cohen et Verlaine, 2013).

Impossible de vérifier la qualité de chaque archive, de choisir précisément au quotidien, même pour une collecte ciblée (par exemple dédiée aux jeux Olympiques ou à des élections), le

17. La définition de l'archéologie des médias, qui apparaît au milieu des années 1990 et interroge les temporalités et matérialités des médias, est débattue. Voir à ce propos : http://pamal.org/wiki/Archéologie_des_média.

contenu qui remontera au cours d'un processus « qui devient de plus en plus automatisé tant au niveau de l'indexation, de la conservation ou de la consultation » (Chaimbault, 2008).

Ces éléments impactent les métiers des archives comme des bibliothèques :

> « Ces évolutions impliquent la définition de nouvelles compétences et de nouveaux profils de postes : par exemple, des "opérateurs numériques" capables d'exploiter au quotidien les processus automatisés de collecte et de traitement, mais aussi des experts en mesure de superviser l'indexation à grande échelle des contenus et de gérer les risques propres à la préservation pérenne des documents numériques alors que les formats et les dispositifs de consultation évoluent et disparaissent très vite. » (Game et Illien, 2006)

Les adaptations ont été rapides comme le montre le récit vivant qu'en livre ci-dessous Gildas Illien, alors conservateur en chef du service du dépôt légal numérique de la BnF, ainsi que responsable technique et trésorier de l'IIPC.

« Les pionniers commencent à moissonner la Toile, généralement à titre expérimental, et saturent, dans l'euphorie des commencements, leurs premiers serveurs de test. Internet Archive, installée dans une petite maison en bois du parc du Presidio, à San Francisco, accueille en stage de jeunes ingénieurs fraîchement recrutés par les BN [bibliothèques nationales] d'Islande, du Danemark, de France ou d'Australie. Ceux-ci reviennent chez eux avec des photos où on les voit boire des sodas et manger des pizzas tout en scrutant joyeusement des lignes de code et d'URL sur des écrans. Dans une ancienne mine du cercle polaire, à Mo i Rana, les Norvégiens installent leur première ferme de serveurs et partent à l'assaut de leur domaine national, le.no. En Islande, un

ingénieur de 25 ans capture et indexe à lui seul tout le Web national, mais ne fait pas cela à temps plein. On apprend sur le tas, on parle de données plutôt que de collections. Les choses se font en masse et à la louche. Les partenaires de l'IIPC sont peu nombreux à proposer une consultation publique de ce qui s'apparente encore à une boîte noire. L'urgence est alors de collecter, l'accès et la conservation de long terme ne sont pas identifiés comme des besoins immédiats. Si bien qu'il n'est pas rare de perdre ou de détruire des données qui, faute de loi, ne sont pas encore devenues inaliénables. Cette époque, profondément sympathique et créative, signe la rencontre du troisième type entre les cadres de bibliothèques nationales multicentenaires et des ingénieurs fous. […]
Mais, début 2010, l'histoire du Web semble s'accélérer, poussant les institutions à élargir sans plus attendre les frontières de leurs interventions patrimoniales. […]
Au même moment, la Bibliothèque du Congrès, la BnF et Internet Archive réalisent ensemble la collecte d'urgence d'un ensemble de sites relatifs au séisme en Haïti. Un an plus tard, elles renouvellent cette coopération spontanée, d'abord pour archiver les sites de WikiLeaks, puis, très récemment, à l'occasion de la révolution du Jasmin en Tunisie et dans le reste de l'Afrique du Nord. Au risque de s'écarter de leurs missions initiales, elles laissent leurs robots s'aventurer dans des zones grises, sans territoire fixe. Car les bibliothèques du consortium ne peuvent plus ignorer des événements et des contenus numériques particulièrement volatils documentant une future histoire du monde qui n'est pas réductible à la somme de leurs histoires nationales. […] »

Illien Gildas, « Une histoire politique de l'archivage du Web », *Bulletin des bibliothèques de France (BBF)*, 2011, n° 2, p. 60-68. Disponible en ligne : http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012. ISSN 1292-8399.

La possibilité de jouer de l'interactivité et de l'hypertextualité des archives les rend également spéciales. Même si le parcours au sein de collections d'archives numérisées ou papier n'est pas forcément linéaire, cette spécificité liée au Web est notable. Comme le rappellent Latzko-Toth et Proulx (in Barats, 2013), il faut prendre en compte les qualités documentaires de l'information en réseau, en termes de « recherchabilité », d'ubiquité, de persistance, de mutabilité et d'« invérifiabilité ». Si la « recherchabilité » de l'information lui permet d'être trouvée par les moteurs de recherche ou de collecte, ce qui détermine l'accès à des données autrement pas ou peu visibles, une partie du Web échappe à la collecte, quand il est mal ou peu indexé, et ce volontairement ou non ; l'ubiquité d'une information copiable et diffusable pose aussi des défis : faut-il conserver la même vidéo qui aurait été postée sur plusieurs plateformes vidéo et apparaîtrait sur YouTube et Dailymotion ? Le mouvement paradoxal de persistance comme de mutabilité rend les contenus à la fois instables et se double de la difficile vérifiabilité des acteurs (en cause l'anonymat et le pseudonymat, mais aussi la masse documentaire qui rend complexe un traitement fin, par exemple dans le cas de la collecte par l'Ina de 20 millions de tweets à la suite des événements du Bataclan, etc.).

Malgré des singularités et des nouveautés, bien des questions que les archivistes ont dû auparavant affronter restent d'actualité. Par exemple la pratique des doublons, fréquents dans les archives du Web, n'est pas inconnue des services d'archives ; de même les collectes d'urgence – à l'instar de celles effectuées au moment des attentats de 2015 par l'Ina et la BnF – ne relèvent pas d'une spécificité liée à l'éphémère du Web, même s'il peut contribuer à en réactualiser les enjeux. En outre, d'autres éphémères, matériels cette fois – tels les messages de réaction aux attentats, de commémorations ou encore les offrandes aux victimes déposées dans plusieurs villes de France – ont été collectés par le passé, notamment par des archives municipales (Bazin, 2017). La collecte des éphémères ne commence donc pas avec le patrimoine nativement numérique :

> « Ce principe de constitution de sources primaires n'est pas, explique Clément Oury, pour [la BnF] une

> nouveauté : ses agents recueillent depuis le XIX^e^ siècle le matériel de propagande électorale (tracts, affiches). » (Oury in Cohen et Verlaine, 2013)

Les chercheurs retrouvent aussi des problématiques connues qui touchent autant à la question de l'authenticité que de l'auctorialité par exemple, car bien des sites sont le résultat de productions souvent collectives, parfois externalisées, etc. Plus généralement, les archives du Web rendent complexes la critique interne, mais aussi externe des documents.

Or, pour comprendre pourquoi, et ainsi rendre ces archives exploitables, le chercheur ne peut faire l'économie de la compréhension de la fabrique de l'archive.

OÙ COMMENCE ET S'ARRÊTE L'ARCHIVE ?

La plupart des institutions de collecte des archives du Web livrent en ligne un aperçu des périmètres et choix de collecte, à l'instar de la BnF[1] qui distingue des collectes larges et des collectes ciblées. Par ailleurs, les chercheurs ont le souci d'essayer de documenter ces sélections et leurs évolutions, que ce soit en ouvrant les boîtes noires de l'archivage (Schafer, Musiani et Borelli, 2016) ou en suivant les traces visibles que ces archives livrent (voir Ben-David et Amram, 2018, sur le Web archivé nord-coréen).

En effet non seulement les institutions, quand elles s'inscrivent dans un cadre juridique fixé, doivent faire porter leurs efforts sur un périmètre défini de sites web, mais aussi mettre en place une stratégie de collecte (en termes de récurrence, de profondeur de l'archivage des sites, de participation ou pas des internautes, etc.) qui va avoir un impact direct sur la représentativité de ces archives. En outre, des barrières à l'archivage peuvent apparaître, notamment pour des raisons techniques (*captcha*, mots de passe), tandis que les réseaux socionumériques, qui feront plus tard l'objet d'un éclairage spécifique, renouvellent les questions de sélection et de capture. Autant d'éléments à découvrir dans cette partie pour tracer les contours de l'archive, qui peuvent varier d'une organisation à une autre, d'un site à l'autre, d'un réseau socionumérique (RSN) à l'autre…

Des archivages en constante évolution

Une archive du Web est loin d'être un objet statique[2] : elle évolue sous l'effet des modalités de collecte, de la profondeur de

1. Voir sur le site de la BnF : http://www.bnf.fr/fr/professionnels/archivage_web_bnf/a.dlweb_collecte_acces_libre.html.
2. Cette section reprend des éléments de Schafer, Musiani et Borelli, 2016.

l'exploration, ainsi que des changements techniques – et, bien sûr, des modèles et paradigmes qui sous-tendent l'archivage.

Lors de l'assemblée générale de l'International Internet Preservation Consortium (IIPC) de 2014, Louise Merzeau soulignait à quel point, malgré l'histoire jusqu'ici brève de l'archivage du Web, on avait déjà pu assister à plusieurs changements aux conséquences de taille pour les archives. Au cours des années 1990, avec la naissance d'Internet Archive, l'archivage du Web suivait un « modèle documentaire » dont l'objectif était un archivage universel, inspiré par les modèles traditionnels et tout particulièrement celui de la bibliothèque. Ensuite, au début des années 2000, ce modèle fut brièvement remplacé par une logique davantage tournée vers les enjeux de mémoire. Une troisième phase mit l'accent sur les aspects de préservation systématique, une sorte de « congélation » à un instant T qui consistait à sauvegarder chaque élément du corpus, pièce par pièce, en un archivage qui, à défaut d'être exhaustif, se voulait représentatif. Enfin, depuis la fin des années 2000, les archives du Web sont construites selon une logique d'« archive temporelle », qui cherche à capturer entièrement l'instabilité du Web – en développant des méthodes d'archivage dynamiques, tout comme le Web est dynamique. L'instabilité, qui avait été considérée comme un dysfonctionnement contingent à l'objet, est de plus en plus perçue comme une de ses caractéristiques essentielles :

> « Paradoxalement, l'instabilité qui caractérise les flux d'information ne constitue donc pas un obstacle à leur mémorisation, mais plutôt une condition, entraînant de nouvelles procédures de sédimentation mémorielle. Parce qu'ils sont instables, les contenus doivent être dédoublés par une information sur l'information, qui anticipe, optimise et instruit leur mobilisation. Les métadonnées désormais associées à tout message ne décrivent pas seulement les énoncés : elles en permettent la segmentation, la distribution et la recomposition, chaque fragment du flux devenant une mémoire activable à volonté, pointant vers d'autres fragments. » (Merzeau, 2012)

RÉCOLTER LES MÉTADONNÉES

« Parmi les éléments de la collecte des documents, il convient de ne pas oublier de récolter les informations sur les pages web, à savoir ce qu'on appelle les métadonnées des documents. Une métadonnée est littéralement une donnée sur une donnée ; plus précisément, c'est un ensemble structuré d'informations décrivant une ressource quelconque. Le recueil des métadonnées doit pouvoir fournir des données sur le contexte technique et historique de la collecte d'une part et du document d'autre part. Les métadonnées fournissent ainsi des renseignements sur le nom du document, sa date de création, de mise à jour, son environnement technique, celui nécessaire pour lire le document (standards d'encodage), leur compatibilité (les standards – les protocoles évoluant, il conviendra d'assurer des migrations régulières en termes de supports de stockage, de langages ou de formats) ; la composition de la page (texte, image, son...), des informations juridiques, etc. »

Chaimbault, 2008 : http://www.enssib.fr/bibliotheque-numerique/documents/1730-l-archivage-du-web.pdf.

Avec cette attention particulière prêtée aux variations du Web « vivant », le Web archivé s'éloigne progressivement de l'idée d'une restitution et permet, comme le pointe Louise Merzeau, de passer d'un fragment à l'autre sans être contraint notamment par la chronologie des flux. Il nécessite donc une compréhension de plus en plus fine des coulisses du stockage et de la circulation des flux d'information (Merzeau, 2014).

Le chercheur Niels Ole Finneman (2015), plaçant au cœur de ses travaux ces questions de temporalité et d'intelligibilité, remarque que tous les corpus d'archives web répondent à trois dimensions temporelles : le contenu original, son accumulation et ses transformations, et enfin l'exploration de l'archive par le spécialiste. Ce dernier devient partie intégrante de

l'intelligibilité des contenus : inscrit dans sa propre époque, il peut introduire des biais, contribuant ainsi à une lecture nostalgique ou présentiste (Schafer, 2015).

Comme le souligne Niels Brügger (2012), un autre aspect très important réside dans le fait que le processus d'archivage du Web crée une série de versions uniques d'un contenu : on n'est presque jamais en train de, tout simplement, « faire une copie ». Des éléments peuvent être perdus (par exemple une image, un bandeau) et autre chose, qui n'était pas en ligne à cet instant T, peut être archivé avec le contenu (par exemple un calendrier anachronique, récupéré d'une page antérieure[3]). Ce qui peut rendre complexe de savoir avec certitude à quoi ressemblait effectivement une partie du Web en ligne à un moment spécifique : chaque archive web est une reconstruction (Ankerson, 2015b).

Plusieurs raisons concourent à expliquer ce phénomène. La première est la profondeur de la collecte et de la capture. Très souvent, les sites web ne sont archivés que partiellement, car le robot *crawler* est programmé pour les capturer seulement à profondeur de quelques clics. Les utilisateurs se trouvent régulièrement face à des pages web manquantes ou non trouvées, mais l'effort porte sur la volonté de capturer des échantillons vastes et représentatifs du Web contemporain dans sa diversité, malgré la « superficialité » que cela entraîne. Par exemple, en France, les collectes larges de la BnF privilégient la quantité ; or, si les 4 millions et demi de sites web collectés dans une année avec ce système sont très rarement préservés dans leur intégralité, c'est aussi le cas de leurs pages web qui sont souvent incomplètes ; des éléments tels que les publicités, les *pop-up* et les bannières sont souvent bloqués avant la collecte. Cela entraîne l'omission d'une partie intéressante et importante du patrimoine nativement numérique, avec laquelle les utilisateurs du Web ont fréquemment un rapport problématique, voire conflictuel, mais qui reste une illustration importante des modèles d'affaires et des stratégies de communication des firmes numériques, basés sur l'économie de l'attention (Kessous, 2012).

3. Cela explique certaines inconsistances qui peuvent surgir lorsqu'on navigue dans le Web archivé – par exemple, quand un widget « calendrier » montre une date différente par rapport à la date de collecte de la page web.

Les polices et caractères peuvent aussi différer dans les archives du Web par rapport aux pages originelles ; si au moment de l'archivage la police d'une page web n'était pas inscrite explicitement dans son code source originel, mais plutôt utilisée par défaut, ce sont les paramètres établis par défaut par le navigateur dans sa version actuelle qui figurent sur la page archivée.

Enfin, la collecte et la sauvegarde des images peuvent poser problème dans ce paysage mouvant : plusieurs pages web des années 1990, désormais archivées, montrent des trous béants là où étaient autrefois leurs images. La raison de ce phénomène est à rechercher autant dans la difficulté technique de la capture, que dans « l'impatience » des robots et dans les objectifs de la collecte à l'époque : Internet Archive était liée à l'entreprise Alexa de Brewster Kahle, une firme qui avait pour objectif de classifier et d'indexer les sites web plutôt que de préserver les images. Aujourd'hui toutefois, afin d'éviter les doublons, ces dernières ne sont pas systématiquement recollectées.

Le chercheur doit donc prendre en compte ces aspects : l'archive du Web n'est pas une copie parfaite de l'état de la Toile, ou même de la page, à un instant T (Brügger, 2012b ; Schafer, Musiani et Borelli, 2016). Certains contenus d'une page ne sont pas forcément archivés (les publicités ou les commentaires par exemple[4]), d'autres ont été récupérés de versions antérieures (logos, calendrier) : il faut considérer la page moins comme une unité qu'un ensemble d'éléments, qui peuvent être collectés séparément :

> « Si l'on considère ainsi qu'en moyenne, une page web contient une quinzaine de liens vers d'autres pages, et environ cinq objets d'origines diverses (sons, images, code, films…), la description technique d'une page demeure ambiguë et floue. » (Chaimbault, 2008)

4. Ainsi, depuis 2010, l'outil UGC et une plateforme de captation des vidéos ont été développés à l'Ina pour archiver les vidéos présentes par exemple sur YouTube et Dailymotion. Mais les commentaires échappent (pour l'instant) à la collecte.

En outre, les pages sont reliées les unes aux autres par des reconstitutions de liens hypertextuels qui peuvent introduire des sauts temporels entre deux pages archivées à des dates différentes, etc. Comparant l'archive du Web à une « archive traditionnelle », Bruno Bachimont peut ainsi noter :

> « Pour une archive traditionnelle, l'enjeu est de conserver un document comme produit d'une activité donnée, dont il est alors une trace probatoire, permettant de renseigner sur la nature de l'activité, de prouver les événements associés. Il est donc essentiel, pour entamer son exploitation, de s'assurer que le document est bien le "bon", c'est-à-dire qu'il est bien ce qu'il prétend être : il doit être "authentique". […] L'authenticité repose sur l'intégrité.
>
> Pour une archive du Web, ce raisonnement ne peut plus tenir. En effet, l'archive du web n'est pas le web, l'archive d'un site n'est pas le site archivé. La raison essentielle tient à la nature même des contenus et des procédures de collecte : en particulier, la durée de captation étant supérieure au rythme de mise à jour du site, l'archive résultant de la collecte rassemble en fait des parties de site renvoyant à des temps ou époques différents du site : une partie correspondant au site au temps t^0, une autre au temps t^1 après une mise à jour, etc. Bref, le site archivé n'a jamais existé comme tel dans le Web. » (Bachimont, 2017a)

Des méthodes alternatives émergent pour la recherche. Les *digital forensics*, ainsi, s'intéressent à la reconstitution de documents critiques à travers les données de navigation, les courriers électroniques, l'historique des recherches, etc. (Kirschenbaum *et al.*, 2010). La diplomatique numérique, elle, propose de contextualiser la valeur du document (Chabin, 2012). Toutes deux viennent tenter de répondre aux interrogations traditionnelles que ces archives numériques renouvellent : comment dater, authentifier un document, combler les lacunes, retrouver le contexte, équilibrer les caractères externes (matériels)

et internes (cohérence des textes) des sources, ou encore évaluer le rapport entre échantillon et tout, singularité et représentativité.

Le recours à la philologie que suggère Niels Brügger, pour comparer les différentes versions d'une page web, témoigne également de ce que les recherches ne s'orientent pas forcément vers des méthodologies en rupture, mais peuvent faire appel à des pratiques antérieures, tout en invitant à les renouveler, les adapter :

> « C'est un déplacement considérable auquel nous assistons. Il nous faut donc inventer une nouvelle herméneutique, celle de la trace collectée, herméneutique à laquelle nous sommes fort peu préparés. Éduqués en maîtres du soupçon pour établir l'authenticité, nous sommes peu versés dans l'art d'exploiter des archives qui sont par essence fautives et incomplètes mais néanmoins fiables et exploitables […] » (Bachimont, 2017a)

Le périmètre de l'archive du Web

Le regard que l'on porte sur l'archive, dans une certaine mesure, définit son périmètre. C'est le cas pour le regard des chercheurs, l'un des premiers publics d'usagers de l'archive du Web. L'analyse de sites web a donné lieu à de riches réflexions méthodologiques et épistémologiques (voir par exemple Barats, 2013), mais qui ont tendu à effleurer la question de l'archive du Web sans, jusqu'à récemment, la prendre en charge frontalement. Niels Brügger a lancé une nouvelle dynamique en 2009, en dessinant les contours d'un usage de l'archive web par les chercheurs (Brügger, 2009 ; 2011) à partir d'éléments distincts : l'objet web (par exemple une image insérée dans une page web), la page web, le site web, la sphère web (un ensemble de pages web liées par une thématique), le Web dans son ensemble (ses normes, ses standards, ses institutions, ses technologies, etc.). Ainsi, les différents niveaux,

formats et éléments documentaires concernés par l'archivage (textes, images, sons, vidéos, graphismes, bases de données, logiciels, codes…) entrent dans un périmètre plus ou moins cohérent selon la manière dont on les analyse.

Toutefois, le regard du chercheur est cadré, bien que non limité, par les dispositifs mis en place par les professionnels de l'archivage numérique en général et du Web en particulier. Jinfang Niu a proposé dès 2012 une vue d'ensemble des enjeux de l'archivage du Web, défini comme le « processus de récolte et de stockage de données enregistrées sur le World Wide Web, de leur conservation sous la forme d'une archive, et de leur mise en accessibilité pour des recherches futures » (Niu, 2012).

Pour Niu, ce périmètre peut être décrit par les processus de travail de cet archivage, qui passent par :

- l'évaluation et la sélection, qui même dans le cas de collections non discriminantes des contenus se font forcément sur la base de critères. Par exemple, pour Internet Archive qui a priori ne trie pas sa récolte, c'est essentiellement le « Web de surface » (indexé par les moteurs de recherche) qui est concerné. Les collections institutionnelles sont plus sélectives, sur la base de critères géographiques, thématiques, événementiels (comme dans le cas des périodes électorales, ou des crises terroristes), ou encore génériques (selon le type ou le format de média). Cette sélection est plus ou moins automatisée ou manuelle, plus ou moins programmée à l'avance ou ouverte à l'intervention (formulaires d'enregistrement, recommandations…). L'évaluation de la valeur peut reposer sur des méthodes très différentes : alors que la NARA (National Archives and Records Administration) américaine évalue la valeur d'un site individuel, la BnF préfère la représentativité (toutes les pages web françaises sans distinction de qualité), et le service des archives web de l'université nationale de Taïwan a recours à l'échantillonnage ;
- l'acquisition : si la tradition institutionnelle de dons et de dépôts est toujours d'actualité, l'archivage du Web a donné lieu à des méthodes originales, comme l'indexation de réseau (*crawling*) qui récolte les contenus par le biais du suivi d'hyperliens. La question des permissions se pose à cette étape, sauf en cas de mandat gouvernemental (en particulier le dépôt légal,

comme en France, en Nouvelle-Zélande, aux États-Unis ou encore au Royaume-Uni) ou de mise en place de clauses de retrait (solutions *opt out*, comme chez Internet Archive) ;
- l'organisation et le stockage : ceux-ci doivent préserver l'intégrité du contenu, en donnant des informations sur l'origine (de la source de l'enregistrement à son adresse en tant que document vivant) et l'ordonnancement (l'agencement au sein de la structure des archives) ;
- la description : les métadonnées décrivant les archives sont générées automatiquement lors de l'indexation (par exemple la signature temporelle de la récolte, la taille, le format, etc.) ou bien induites à partir d'une extraction des métadonnées du code des pages d'origine ;
- l'accès et l'utilisation : ils sont déterminés par le contexte légal de l'archive du Web, avec une tendance à la restriction sur le modèle des « *dark archives* », qu'on ne peut consulter qu'in situ « à l'ombre » des bibliothèques, par opposition aux archives ouvertes (Smit, van der Hoeven et Giaretta, 2011). Les potentialités de la recherche reposent sur la richesse des métadonnées de description, des outils d'indexation et des choix d'interface.

Pour les professionnels, le cahier des charges d'un projet d'archivage du Web résume ces problématiques en cinq recommandations formulées par l'IIPC Preservation Working Group : la mise en place d'objectifs à but juridique et/ou scientifique ; l'évaluation des possibilités et contraintes légales ; l'approche raisonnée de la création de collections selon des critères ; l'identification des problèmes de mise en collection (techniques et organisationnels) ; la stratégie de conservation à long terme (métadonnées, formats…).

De nombreuses contraintes limitent le périmètre des archives du Web, notamment pour les institutions contraintes par le droit d'auteur. Internet Archive, qui prône une politique de numérisation massive, revendique une responsabilité civique dans l'accessibilité publique aux contenus, quitte à contourner ce que la fondation considère comme des barrières fixées par l'économie et le droit de l'édition et des archives, par exemple l'application de mesures techniques de

protection du droit d'auteur trop contraignantes – telles que les DRM[5]. Le périmètre de ses archives en est d'autant plus élargi, avec une ambition non départie d'idéaux universalistes (Paloque-Bergès, 2014). C'est aussi l'approche de beaucoup d'organisations non institutionnelles, fondations privées, jeunes entreprises ou initiatives individuelles, qui étendent le périmètre de l'archive du Web aux activités culturelles sur Internet, dans une logique d'auto-archivage des productions individuelles. Par exemple, le Google Cultural Institute produit des outils accompagnant les utilisateurs dans la création de galeries de vie numérique sur leurs sites web personnels. Récusant le vocabulaire des professionnels du patrimoine, comme « commissaire d'exposition numérique », il encourage le « mariage du professionnel et de l'amateur » dans le domaine de la conservation numérique. Ces approches exogènes aux institutions du patrimoine invitent à interroger la manière dont le numérique altère la perception de ce qu'est un document, une archive, ou encore une collection, au sens technique, mais aussi culturel et social. Concernant les contraintes limitant le périmètre de l'archivage, les collections de blogs ont aussi retenu l'attention, de par les problèmes qu'ils posent en termes de droit d'auteur et de la personne, de responsabilité d'hébergement, de filtrage et d'éditorialisation des informations, de frontières floues entre production professionnelle et amateur, de limites labiles entre contenu d'auteur et commentaires du public, etc. Des projets spécifiques ont été mis en place pour les prendre en charge, comme BlogForever, projet collaboratif collectant, conservant, administrant et réutilisant des archives de blogs, financé par la Commission européenne[6].

Il apparaît donc, comme le rappellent Sarah Atkinson et Sarah Whatley (2015), que les archives numériques doivent être mises en perspective avec l'espace public numérique. L'utilisateur et le public jouent un rôle dans la construction du périmètre de l'archive, favorisant les pratiques de l'archivage collaboratif et ouvert.

5. Digital Rights Management (gestion des droits numériques).
6. Pour en savoir plus, consulter : https://cordis.europa.eu/project/rcn/98063_fr.html.

L'archivage des réseaux socionumériques, quelles spécificités ?

Si l'archivage du Web a bénéficié de l'initiative précoce de Brewster Kahle, le paysage numérique et ses usages ont profondément changé depuis 1996, notamment avec l'arrivée des réseaux socionumériques (RSN), fondés sur des dispositifs de flux. Ainsi Frédéric Clavert (2018a) note à propos de Twitter que « collecter des tweets, notamment, via une API, c'est transformer un flux constant en archive figée. La notion de source, flux originel intarissable, n'a jamais été une métaphore aussi actuelle ». Les RSN proposent par ailleurs des modalités de participation et d'accès, qui peuvent rendre l'archivage complexe : identifiants et mots de passe, statuts privés ou semi-publics des contenus, usages de protocoles spécifiques, notamment concernant les vidéos, encapsulage de liens contenant des URLs parfois réduites, etc. Les contenus des RSN ne sont donc pas toujours aisément accessibles ou/et faciles à collecter, sans compter les changements de protocoles ou de politiques utilisateurs qu'ils introduisent fréquemment. Comme le rappelait Annick Le Follic, alors chargée de collections numériques au département du dépôt légal de la BnF, dans un entretien le 21 mars 2016 :

> « La limite de notre archivage des réseaux sociaux est technique : ces plateformes changent souvent de technologies et de paramètres, donc il nous faut donner à chaque fois une instruction manuelle à Heritrix[7] pour qu'il capture bien les contenus qui nous intéressent. En particulier, les protocoles https[8] nous posent parfois des problèmes, tout comme Facebook lorsqu'il utilisait des "captcha"[9]. »

7. Robot d'indexation utilisé par la BnF mais aussi par Internet Archive : https://webarchive.jira.com/wiki/spaces/Heritrix.
8. Protocole web sécurisé.
9. Entretien mené par M. Borelli et V. Schafer dans le cadre du projet ASAP, 21 mars 2016 : https://asap.hypotheses.org/168.

Les RSN n'en demeurent pas moins des témoins et supports de nos vies numériques, qui ne pouvaient rester en dehors de la réflexion sur l'archivage du Web.

La Bibliothèque du Congrès (LoC) aux États-Unis a ainsi passé un accord en 2010 avec l'entreprise Twitter pour récupérer tous les tweets émis depuis 2006 et poursuivre cette conservation. Reste qu'à ce jour cette collection n'est pas encore accessible pour les chercheurs et soulève diverses questions, amenant même la LoC à revenir sur son projet d'exhaustivité pour se concentrer sur un périmètre plus restreint et sélectif de collecte[10]. En effet, les outils disponibles pour faire des recherches dans ces fonds gigantesques sont un enjeu majeur (le nombre de tweets journalier est passé selon la LoC de 140 millions début février 2010 à 500 millions par jour en octobre 2012). Dans un document de janvier 2013, intitulé « Update on the Twitter Archive At the Library of Congress[11] », la bibliothèque notait ainsi que réaliser une recherche sur la période 2006-2010 pouvait prendre 24 heures, et elle faisait le constat que les technologies disponibles pour accéder à ces données n'étaient pas encore aussi avancées que celles permettant de les collecter.

Bien sûr l'accord entre la bibliothèque étasunienne et l'entreprise Twitter pose également la question des modalités concrètes d'accès à ces archives : leur accessibilité pour des chercheurs par exemple européens impliquera-t-elle de devoir venir à la LoC ?

Des initiatives européennes ont aussi été engagées, mais avec des périmètres plus restreints, appuyés par exemple en France sur le cadre du dépôt légal du Web. La collecte de Twitter par la BnF et l'Ina apporte des éléments complémentaires à une réflexion sur le patrimoine des RSN.

Tout d'abord, si la BnF et l'Ina archivent une partie de Twitter, elles n'ignorent pas les autres RSN, mais peuvent

10. Voir l'article de *The Verge* du 26 décembre 2017, « The Library of Congress will no longer archive every tweet » : https://www.theverge.com/2017/12/26/16819748/library-of-congress-twitter-archive-project-stalled.
11. Library of Congress, « Update on the Twitter Archive at the Library of Congress », décembre 2017 : https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

rencontrer plus de difficultés pour les collecter. Les deux institutions ont davantage archivé Twitter que Facebook par exemple, car les contenus de Facebook ne sont pas tous publics, outre les difficultés techniques précédemment évoquées. Et pourtant les Français sont davantage présents sur Facebook et la diversité sociologique y est mieux représentée[12].

Comme pour le Web, le périmètre de collecte est aussi sélectif pour les RSN. Si l'Ina a pris la mesure de l'intérêt de l'archivage de Twitter et lancé des collectes dès 2014, l'équipe dédiée au DL Web le fait dans le cadre de son périmètre lié à l'audiovisuel : elle suit ainsi les comptes d'acteurs clés du monde audiovisuel français, soit environ 13 000 utilisateurs et 400 *hashtags*.

Son expérience s'est aussi manifestée lors des attentats de 2015, au moment où des millions de tweets ont réagi aux événements autour de *Charlie Hebdo* puis à ceux de novembre 2015 (suscitant aussi la réactivité des chercheurs qui ont également très rapidement lancé des collectes de ces tweets[13]).

Comme le note Zeynep Pehlivan (DL Web Ina) qui revient sur cet archivage réalisé en urgence :

> « Nous avons poursuivi les collectes sur les attentats après 2015, par exemple à Nice à l'été 2016. Nous avons aussi des archives relevant d'attentats qui ont eu lieu en Europe, à Bruxelles, Londres ou Manchester. En effet s'ils ne se sont pas passés en France, ils ont été profondément relayés par les médias français et sont entrés rapidement dans les *trends* [principales tendances de mots-clés sur Twitter] de Twitter, car les Français ont réagi. Ces tweets font partie intégrante du contexte médiatique et permettent en outre au chercheur de mettre en perspective les tweets de notre cœur de corpus du dépôt légal. Par contre, on

12. Pour un aperçu des chiffres, voir : https://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2018/.
13. C'est le cas de la collecte de Romain Badouard qui sert de base à sa réflexion sur le « Je ne suis pas Charlie » (Badouard, 2016), de celle du canadien Nick Ruest, dont les données sont accessibles en ligne, ou encore de celles de Giglietto et Lee (2015).

> ne fait pas des collectes pour tous les attentats dans le monde, seulement pour ceux qui ont un écho fort en France, en particulier dans le monde de l'audiovisuel, qui est notre périmètre dans le cadre du dépôt légal du Web[14]. »

L'Ina a pleinement conscience de l'intérêt de démarrer tôt la collecte, de ne pas rater le pic de tweets ou la montée d'un « mot-dièse » (des mots-clés précédés d'un signe #, appelé *hashtag*, permettant d'étiqueter les tweets).

> « Or le service est fermé la nuit ou le week-end. Aussi nous avons décidé d'archiver dorénavant automatiquement les principaux *trends* en France. Nous avons ainsi une veille automatique complémentaire, même en dehors des heures d'ouverture, sur des mots-dièses qui montent et sont en général portés ou repris dans les médias. Aujourd'hui les journalistes aussi participent et suivent en effet Twitter et ces mouvements[15] », ajoute Zeynep Pehlivan.

Si l'aspect des archives du Web peut changer d'une institution à une autre, le cas de Twitter est particulièrement révélateur, comme nous l'avons mentionné en introduction : la BnF utilise le robot de capture Heritrix et obtient des résultats proches d'une capture d'écran, tandis que l'Ina passe par l'API (interface de programmation) publique de Twitter et ne capte pas les images de fond. Il est possible de récupérer a posteriori les données de Twitter de façon payante : les deux interfaces de programmation, API Search et Streaming par lesquelles passe l'Ina, sont gratuites et publiques. La première permet à un utilisateur de remonter à un contenu particulier sur les sept derniers jours, tandis que la seconde permet de capter un flux au fur et à mesure pour une requête précise. Mais l'API publique a des limites : on ne

14. Entretien réalisé par Valérie Schafer fin 2017 dans le cadre d'un article dédié au patrimoine nativement numérique des attentats en Europe pour un dossier de la *Gazette des archives* (nº 250) coordonné par Maëlle Bazin et Marie van Eeckenrode.
15. *Ibid.*

peut collecter plus de 1 % du total des tweets émis au plan mondial à un instant T. Cette limite a notamment été dépassée au moment du pic de flux lié aux attentats parisiens, et même les 20 millions de tweets conservés par l'Ina sur les événements du Bataclan ne constituent donc pas une collecte exhaustive de ce qui s'est dit sur Twitter autour du 13 novembre 2015. Ajoutons que la collecte dépend des mots-dièses sélectionnés et que certains peuvent échapper à l'archivage qui se fait en urgence. D'autres biais ou limites ne peuvent être ignorés du chercheur : par exemple le nombre de retweets (republication de tweets par un autre usager) d'un message s'arrête à la date de l'archivage du tweet, impliquant donc de sérieuses précautions sur l'interprétation de cette donnée.

Reste qu'au-delà de ces limites, le volume archivé au moment des attentats parisiens est tel qu'il peut être considéré comme représentatif, à défaut d'être exhaustif, d'autant que l'Ina s'applique à documenter sa collecte en intégrant notamment des informations sur les données manquantes, en archivant les messages signalant une restriction dans la collecte, etc. Évidemment, il faut souligner une autre limite à la représentativité, mais qui ne dépend pas de la collecte : les publics de ces plateformes sont spécifiques « comme le sont les lecteurs de journaux ou les tenants de la conversation de bistrot. Mais ces traces peuvent sous certaines conditions donner accès à certains processus qu'on ne pouvait chiffrer jusqu'ici » (Boullier, 2015).

Les barrières, limites, verrous à l'archivage

Déjà évoquées, la disparition des pages web, la volatilité des contenus et l'évolution générale des réseaux sont les limites fondamentales rencontrées par l'archivage du Web. En 2013, la durée de vie moyenne d'une URL est de 9,3 ans ; celles qui ne survivent pas entretiennent le « *link rot* » (la décomposition des liens). Un « lien mort » est d'autant plus dommageable qu'il a pu servir de référence, voire de garantie institutionnelle, comme en a témoigné l'affaire des articles disparus de la Cour suprême américaine révélée par le *New York Times* en 2013 – on parle alors de « *reference rot* ». Les liens et contenus web s'évanouissent

au gré de la fermeture d'hébergeurs ou de plateformes, de la réorganisation de l'architecture d'un site, ou parce qu'un auteur a tout simplement choisi de supprimer un contenu, voire d'effacer complètement sa présence numérique, ce que l'on surnomme « *infosuicide* ».

Le Web peut également, tout en restant bien vivant, résister à l'archivage. Pour des raisons techniques, tout d'abord, dans la mesure où il peut être difficile pour les dispositifs d'archivage automatique de capturer des contenus et objets mis en forme par des technologies non prises en charge par le dispositif ou obsolètes. Suivant une logique de flux, le Web dynamique tend à encapsuler des contenus hébergés ailleurs, une page n'étant que de plus en plus rarement une unité homogène. Ainsi, ces dispositifs peuvent avoir tendance à reconstituer des pages « à trous ». Par exemple, le langage JavaScript permettant l'encapsulation de contenu a été l'un des premiers obstacles au moissonnage de données web par Heritrix, produisant des archives de pages web qui sont des coquilles vides. L'enchâssement de plusieurs types de logiciels de gestion de contenu et la superposition de plusieurs couches de code peuvent également compliquer la tâche d'une collecte numérique. C'est le cas de la republication ou de l'administration de forums internet, notamment des groupes Usenet : parfois mal gérés par leurs administrateurs, difficiles à naviguer, impossibles à collecter, ils tendent à devenir des « ruines numériques » sur le Web (Paloque-Bergès, 2018 ; 2017).

Des barrières plus proactives peuvent être mises en place par les hébergeurs, les administrateurs et les auteurs. Le problème du verrouillage par mot de passe est un classique, que l'on retrouve de manière généralisée sur les plateformes de réseaux sociaux. Le recours à un code contractuel est également une technique ancienne, comme dans le cas du *robot.txt*, une formule insérée dans le code source d'une page web par son créateur. Cette technique « a pour but principal de permettre à un éditeur d'exclure certains de ses documents du champ d'action des agents logiciels appelés "crawlers" utilisés par les moteurs de recherche pour prendre connaissance des documents » (Sire, 2015, p. 188).

Toutefois, comme l'analyse Guillaume Sire, ce contrat de code repose sur un consensus léonin, c'est-à-dire régi par des

rapports de force déséquilibrés. Google peut choisir de passer outre ce protocole tout comme certaines institutions d'archivage du Web, ces dernières en vertu des modalités du dépôt légal (Niu, 2012).

Pour les archives du Web, comme pour nombre d'autres artefacts techniques qui peuplent l'internet, un certain nombre de barrières, limites et verrous à l'archivage prend forme lorsque l'infrastructure de l'internet, du matériel au logiciel, joue un rôle social et politique dans leur « fabrique », notamment à des niveaux micro et parfois triviaux (Cheniti, 2009). Nous prendrons ici deux exemples qui ont trait aux contributions volontaires des internautes à l'archivage du Web[16].

En janvier 2015, Andrew Bontrager, un utilisateur des services de la fondation américaine Internet Archive, commente un changement sur les conditions d'utilisation :

« *…from your terms of use:*
"…Further, you agree not to recirculate your password to other people."
This is a hardship.
I had previously done this because I didn't realize you had the provision there.
***Sometimes, I want to contribute a large file to the archive, but my internet connection is slow or limited by a data plan. In those instances, I have to give my credentials to another worker so he can do it for me. Thus, I'm asking an exemption[17]* ».**

Et quand l'Archive Team se présente, elle esquisse les

16. Voir aussi https://webcorpora.hypotheses.org/460.
17. « Tiré de vos conditions d'utilisation: "De plus, vous êtes d'accord pour ne pas rediffuser votre mot de passe à des tiers". C'est une grosse contrainte. Je l'avais déjà fait, car je n'avais pas réalisé que vous aviez cette disposition. Parfois, je souhaite ajouter un gros fichier à l'archive, mais ma connexion internet est lente, ou j'ai un barème pour l'échange des données. Dans ces cas, je dois donner mes identifiants à un autre travailleur pour qu'il puisse le faire pour moi. Donc, je demande à être exempté. » (Notre traduction.)

profils et les types de contributions qui lui seraient utiles ainsi :
« *This project is composed of volunteers, currently coordinated by Jason Scott.*
If you're wondering where to stick your nose in, we could use:
Warriors, You will run the Archive Team Warrior on any PC's you have with spare bandwidth. […]
Writers, who can create clear essays and instructions for archivists and concerned parties.
People with Lots of Hosted Disk Space who have a proper hosted webserver and fat pipe, who are willing (when asked) to consider hosting mirrored dead sites or archives. […][18] ».
Deux exemples donc, ayant trait, le premier, à une démarche collaborative de contribution là où les conditions techniques ne permettent pas à l'individu de contribuer seul, le deuxième à une hiérarchie de contributeurs établie sur la base des ressources techniques de stockage et réseautage à leur disposition. Les deux montrent bien comment les contributions voient s'établir des limites non seulement par la volonté et l'organisation humaines, mais également par des facteurs tels que la rapidité d'une connexion internet ou la possibilité d'y accéder de façon constante, la présence de « goulots d'étranglement » qui rendent impossible l'archivage de pages protégées par mot de passe, la capacité à mettre en œuvre une tâche partagée au moyen de différents outils et protocoles et de leur interopérabilité, ou encore la disponibilité de ressources techniques de stockage ou de mémoire et leur ouverture à la communauté.

18. « Ce projet est composé de volontaires, qui sont actuellement coordonnés par Jason Scott. Si vous vous demandez où fourrer votre nez, on aurait besoin de : Guerriers, vous ferez tourner le Guerrier de l'Archive Team sur tout ordinateur à votre disposition qui a de la bande passante non utilisée ; Écrivains, qui peuvent écrire des essais et des instructions clairs pour les archivistes et autres tiers ; Gens avec Beaucoup d'Espace Disque, qui font tourner un web serveur et ont de gros tuyaux, et qui sont disponibles, quand on leur demande, pour héberger des miroirs de sites web qui ne sont plus maintenus, ou des archives. » (Notre traduction.)

« *Link rot* », « *reference rot* », « *infosuicide* », « *digital ruins* » : autant d'images d'un Web en décomposition, dont la logique entre pourtant dans ce que l'archéologie des médias appelle les « médias zombie », où l'information ne meurt jamais tout à fait, car elle survit sous une forme ou une autre (Chun, 2011). De fait, ce dépérissement stimule la résilience. Ainsi, Tim Berners-Lee lui-même a été l'un des promoteurs les plus actifs de techniques de liens pérennes au sein du monde des développeurs web, derrière le slogan « *Cool URIs*[19] *don't change* ».

Des enjeux de gouvernance

En 1980, le philosophe et sociologue Langdon Winner se demandait dans un article qui a fait école : « Est-ce que les artefacts sont politiques ? » (*Do artifacts have politics ?*). Winner pose la question de la neutralité technologique et recherche en se penchant sur les objets techniques les « arrangements de pouvoir et d'autorité dans les associations humaines, ainsi que les activités qui se passent à l'intérieur de ces arrangements » (Winner, 1980, p. 123). Si l'on souhaite appliquer cette hypothèse aux archives du Web, il s'agit de comprendre en quoi dans l'archivage du Web existent des formes spécifiques d'autorité et de pouvoir (Denardis, 2014) qui dessinent une sorte de microcosme de la gouvernance d'Internet[20].

L'archivage du Web repose sur un modèle multi-parties prenantes. Une variété d'acteurs est concernée : des fondations comme Internet Archive ; des organisations transnationales, à commencer par l'IIPC ; la société civile (des membres de l'Archive Team à d'autres initiatives fondées par des communautés de chercheurs) ; et enfin le secteur privé

19. Les URIs (Uniform Resource Identifiers) sont les identifiants qui complètent les URLs (Uniforme Resource Locators) pour la composition et la reconnaissance des pages web.
20. Cette démarche a occupé certains de nos travaux récents (Schafer *et al.*, 2016 ; Musiani et Schafer, 2019) sur lesquels cette section se fonde.

(par exemple, Google, qui s'est impliqué dans la conservation du patrimoine numérique natif en rendant disponible un certain nombre de groupes du forum numérique Usenet ; Paloque-Bergès, 2017). Ainsi, on retrouve dans l'archivage du Web les principales catégories d'acteurs impliqués dans la gouvernance d'Internet, leurs tensions, mais aussi leurs alliances. Des expériences de collaboration entre des institutions d'archivage et des équipes de recherche voient de la sorte régulièrement le jour ; la BnF a par exemple associé notre équipe Web90[21] à une réflexion sur l'implémentation de la recherche en plein texte dans les archives web des années 1990 et, à un niveau plus global, le réseau RESAW[22] associe des chercheurs et des professionnels de l'archivage. Internet Archive va encore plus loin en promouvant explicitement des initiatives *bottom-up* [du bas vers le haut] destinées à revaloriser l'intervention humaine.

L'archivage du Web n'échappe cependant pas à des tensions ayant trait à la standardisation, un des enjeux traditionnellement le plus vif de la gouvernance d'Internet, et à des visions et imaginaires divergents, des communs aux formats propriétaires. Nous avons ainsi évoqué la mission de la BnF, menée dans le respect de la propriété intellectuelle et la protection des données personnelles qui contraste avec la mission « universelle » que s'est assignée l'Archive Team, fondée sur la disponibilité des ressources informatiques et le souhait, de la part des utilisateurs, de les partager. Dans le premier cas, on voit en partie le poids d'un héritage historique et des questions de souveraineté liées au dépôt légal ; et, dans le second, le lien direct entre la capacité technique et l'archivage effectué.

L'archivage du Web révèle également la présence de tensions géopolitiques, illustrées par le blocage d'Internet Archive par la Chine (Kahle, 2014b) ou encore par l'appel de Brewster Kahle, à la suite de la victoire électorale de

21. De 2014 à 2018 ce projet, financé par l'Agence nationale de la recherche et auquel ont contribué les auteurs de l'ouvrage, a exploré l'histoire, la mémoire, le patrimoine du Web des années 1990 en France : https://web90.hypotheses.org.
22. Réseau de recherche européen, RESAW signifie *A Research Infrastructure for the Study of Archived Web Materials*. Il a été établi en 2012 à l'initiative de Niels Brügger : http://resaw.eu/about/.

Donald Trump, à un financement participatif pour créer par précaution une copie complète des collections numériques de l'Internet Archive hors des États-Unis.

On retrouve aussi des dynamiques qui rappellent le problème de la fracture numérique : la présence des pays en voie de développement dans le Web archivé n'est aucunement proportionnelle à leur présence croissante au sein du Web vivant (Gomes *et al.*, 2011). Un certain nombre d'associations régionales se proposent d'épauler l'action globale de l'IIPC et de faire office de « sous-forums » pour coordonner le transfert de compétences pratiques – des initiatives se développent notamment dans le sud-ouest de l'Asie. Cependant, il existe encore des régions du monde qui restent largement « non archivées », en particulier en Inde, en Amérique latine et en Afrique. Comme l'expose la conférence « The Memory of the World in the Digital Age » (Unesco, 2012), parmi les problèmes élémentaires de l'archivage numérique se trouve la simple absence de ressources techniques, légales et financières. Pour pallier le risque de perdre des ressources culturelles, politiques et sociales importantes, des institutions « du Nord » ont entrepris de préserver certaines d'entre elles (par exemple, l'université d'Heidelberg effectue une collecte du Web sociopolitique chinois) ; mais à long terme une réponse durable devra résider dans le développement d'initiatives locales.

On retrouve dans l'archivage du Web la relation complexe entre différentes pratiques et sources d'autorité ou de normativité, de la technologie au marché, de la concertation transnationale et internationale aux standards et aux droits. Cette pluralité a déjà été identifiée pour la gouvernance d'Internet (Bygrave et Bing, 2009 ; Badouard *et al.*, 2013). Le « sauvetage » de Geocities opéré par l'Archive Team suite à la fermeture de la plateforme d'hébergement de pages personnelles par Yahoo!, les collectes d'archives et de données privées par Twitter et Facebook, le dépôt légal dans plusieurs pays, la charte de l'Unesco, l'action de standardisation de l'IIPC : ces différents instruments de gouvernance coexistent et se superposent partiellement. L'archivage du Web réactive donc les mêmes polarisations, négociations et dynamiques qui avaient émergé lors de la naissance de la gouvernance

d'Internet, notamment avec le Sommet mondial sur la société de l'information en 2003 et 2005 (Working Group on Internet Governance, 2005).

COMMENT NAVIGUER DANS L'ARCHIVE ?

L'archive du Web cherche à reproduire l'interactivité qui existait au sein du Web vivant en permettant de cliquer sur les liens et de naviguer dans la Toile. Elle présente toutefois des caractéristiques en termes de temporalités, d'interfaces, de granularité, d'accompagnement des données par des métadonnées, qui rendent explicite le fait que l'archive du Web n'est pas une copie à l'identique du Web au moment de son archivage. Naviguer dans la Toile du passé implique donc des défis et des précautions théoriques comme pratiques, qui interrogent au final la possibilité de repenser ce Web du passé en contexte.

Les temporalités de l'archive du Web

La question des temporalités est probablement l'un des enjeux les plus aigus en matière d'exploitation des corpus conservés. L'archive du Web est instable et signe la « fin de la matérialité documentaire » par le rassemblement de documents « modulables et mobiles », en contradiction avec la vision traditionnelle de l'archivage dont la fonction serait de « figer et [de] stabiliser ». Aussi, l'archive en ligne est marquée du sceau d'une « temporalité brève qui s'accorde mal avec le temps de la recherche historique » (Gebeil, 2016).

S'il ne faut pas minimiser les difficultés posées, il s'agit surtout d'acclimater les pratiques de recherche à de nouveaux régimes de temporalité, dont l'archive hérite du Web lui-même[1]. Serge Noiret nous avertissait en 2011 :

1. Voir l'ouvrage collectif *Temps et temporalités du Web*, Presses universitaires de Paris Ouest, Paris, 2018, issu du colloque éponyme organisé à l'Institut des sciences de la communication du CNRS en décembre 2015.

> « le *digital turn* [tournant numérique] a rendu précaire un certain nombre de concepts chers aux historiens comme celui de la pérennité des sources et la capacité de reproduire dans le temps une analyse qui s'y réfère. » (Noiret, 2011)

Comme Joe Chip, le héros plongé en pleine régression temporelle dans *Ubik* de Philip K. Dick (1969), les utilisateurs de l'archive du Web sont soumis à des régimes chronologiques nouveaux. En premier lieu parce que la sauvegarde d'un site aux mises à jour fréquentes se heurte à l'impossibilité d'une captation totale des données qui le composent : toutes les modifications et ajouts ne peuvent pas être archivés (Mussou, 2012). Ainsi, les archives du site tf1.fr entre 1996 et 2000 dans Internet Archive donnent à voir un corpus réalisé au travers de 18 collectes successives. Pour l'année 1997, ce sont trois captations qui permettent de consulter le site de la première chaîne. À la BnF, les collectes portent sur plusieurs millions de sites archivés depuis 2011 à des fréquences variables, d'« une fois par semaine » à « une fois par an », associées à des « collectes projet » autour d'un sujet particulier[2]. Dans ce cadre, aucune garantie n'existe sur la possibilité de retrouver un site dans son état initial à une date donnée (Brügger, 2012a), chaque état étant le patchwork des modifications intervenues depuis la dernière captation.

Dans le cadre d'une navigation entre les sites, l'archive du Web doit être traitée comme un pavage discontinu de couches temporelles différentes : la page du *Monde* dans la Wayback Machine du 21 février 1999 renvoie par le lien « Nouvelles technologies » à celle du 8 février 1999 (Schafer et Thierry, 2015). L'image du réseau donnée par l'archive est temporellement désaccordée.

À l'échelle de la page et de ses ressources (images, liens, fichiers embarqués divers…), un temps désarticulé est également à l'œuvre : certains contenus d'une page ne sont pas archivés (les publicités ou les commentaires lorsqu'ils sont

2. Voir http://www.bnf.fr/fr/collections_et_services/anx_pres/a.collectes_ciblees_arch_internet.html.

permis par exemple sur les sites de la presse en ligne) ou recollectés, comme évoqué précédemment. Ce dédoublonnage des ressources conduit par exemple à trouver le logo endeuillé de noir du CNRS sur la page d'accueil du site captée en août 2015 par la BnF alors qu'il a été mis en place suite aux attentats de novembre 2015… Des fonctionnalités récemment introduites dans certaines archives du Web peuvent toutefois permettre d'identifier la date de collecte de chaque élément d'une page web archivée par rapport aux autres qui composent cette même page, rendant désormais visibles et explicites ces patchworks temporels[3].

Enfin, la page web archivée elle-même, en tant qu'espace d'affichage ou contenant informationnel, ne comporte pas forcément de date de création, pas de date de modification, mais seulement une date d'archivage, ce qui rend l'analyse diachronique hasardeuse :

> « *there remains a question of the documents' timestamps: The timestamp of the snapshot of a past version of a URL is that of the date of archiving, not necessarily the last updated date of that URL* [...] *To solve this problem, researchers usually aggregate the archived URLs per year, which results in an approximation of an historical hyperlink network with a large margin of error*[4]. » (Ben-David et Huurdeman, 2014)

À la contrainte des régimes de temporalités désaccordés qu'impose l'archivage s'ajoutent les décalages entre les

3. Voir par exemple dans le cas d'Internet Archive, le billet posté sur leur blog le 5 octobre 2017 par Mark Graham "Wayback Machine Playback... now with Timestamps!": https://blog.archive.org/2017/10/05/wayback-machine-playback-now-with-timestamps/.

4. « Une question reste en suspens à propos de l'« horodatage » des documents : l'« horodatage » d'une ancienne version d'une URL archivée est l'« horodatage » qui correspond à son moment d'archivage, pas nécessairement celui de la dernière mise à jour de cette URL [...] Pour résoudre ce problème, les chercheurs agrègent les URLs archivées par année ce qui crée un réseau de liens avec une large marge d'erreur dans les dates utilisées. » (Notre traduction.) La question de la datation est également sensible dans la thèse de Quentin Lobbé (2018) sous la direction de Pierre Senellart et Dana Diminescu. Se fondant sur la notion de « fragment Web », il explore la possibilité de retrouver sa date d'édition et non la seule date d'archivage.

temporalités *en ligne* et *hors-ligne*. Comme le rappelle Clément Oury dans le domaine des sites politiques, une fois le scrutin achevé, on observe une rapide disparition des pages utilisées pendant la campagne, notamment sous l'effet des recompositions plus ou moins rapides du paysage politique :

> « On a vu, notamment au lendemain du premier tour des élections régionales de 2010, des candidats fermer définitivement leur blog lorsqu'ils ralliaient une liste d'union. » (Oury, 2012)

Pendant la seule campagne pour les élections législatives de 2007, la moitié des sites créés pour l'occasion avait disparu cinq mois plus tard.

À l'inverse, comme le souligne Claude Mussou (Ina), l'archive du Web se constitue au fil de l'eau, à mesure que le corpus s'alimente par sédimentations successives, les collectes s'ajoutant les unes aux autres (Mussou, 2012).

En outre, le hors-ligne pilote en partie l'archive du Web : face aux attentats qui ont frappé la France et en particulier *Charlie Hebdo* en 2015, la BnF comme l'Ina ont choisi de mener des collectes d'urgence.

Les nouveaux régimes de temporalités de l'archive en ligne nous poussent probablement à rompre avec le confort que comporte l'utilisation d'archives datées et précisément identifiées que l'époque contemporaine nous avait habitués à utiliser. Toutefois les collègues spécialistes de périodes plus reculées et moins prolixes en documentation écrite ont déjà affronté des questions semblables. D'un regard vers le passé peut naître une manière d'envisager l'avenir, fut-il numérique.

Interaction et interactivité avec l'archive du Web

L'exploration des archives du Web implique en outre de se soumettre à un régime d'interactivité porté par les interfaces et services qui mettent à disposition du chercheur les masses de données préservées.

Intimement liées à l'esprit du projet initialement conçu par Brewster Kahle à la fin des années 1990, les archives du Web proposent une expérience très proche de celle de la navigation en ligne, progressivement enrichie par de nouvelles fonctions (recherche en plein texte, API diverses, etc.) qui s'adaptent à un enrichissement des corpus, particulièrement avec l'entrée des réseaux socionumériques dans l'orbite de la conservation.

Comme elle avait pesé sur la mise en images et en mots d'Internet, la bibliothèque continue d'être une référence incontournable pour penser l'archive du Web. En 2011 Brewster Kahle rappelait son ambition de faire d'Internet Archive « une bibliothèque numérique » dont la « visée [est] à la fois sociale et technologique » et qui permet un « accès universel à l'ensemble de la connaissance : tous les livres, toute la musique, toutes les vidéos, accessibles partout, par tous » (Kahle, 2014a). C'est cette vision qui l'habite depuis l'origine du projet tel qu'il le décrit en 1997 dans *American Scientific* (Kahle, 1997).

Cette vision explique qu'en 2001, quand naît la Wayback Machine[5] qui permet l'accès aux ressources d'Internet Archive, les sites et leurs pages constituent l'unité de base de la consultation. Internet Archive contient des sites comme une bibliothèque contient des livres.

Encore aujourd'hui, l'entrée principale dans l'archive se fait par l'adresse du site. La navigation dans les versions successivement archivées se fait également à l'échelle du site dans le cadre de ce que Anat Ben-David et Hugo Huurdeman désignent comme une « *single URL approach* [approche par URL unique] » (Ben-David et Huurdeman, 2014).

Bien entendu, une navigation au fil des liens est possible entre les sites archivés, mais sans garantie que les liens aboutissent.

Ce régime d'interaction avec l'archive qui est fondé sur la double métaphore de la bibliothèque et de la toile n'est pas sans poser des problèmes. Le premier d'entre eux, comme le souligne Megan Ankerson, est probablement l'importance disproportionnée donnée au facteur temporel dans une sorte de

5. Pour accéder à la Wayback Machine : https://archive.org/web/.

voyage « chrono-touristique » qui s'impose au chercheur au sein des archives (Ankerson, 2015b). La Wayback Machine ne se prive pas de faire reposer sa communication sur l'invitation à un « voyage dans le passé » mis en avant jusqu'à son interface, surmontée par le slogan « *Explore more than 345 billion web pages saved over time* ».

Ces biais introduits par l'interface et les conditions de collecte des données, les institutions responsables de l'archivage ont tenté de les pallier.

La première étape a consisté à mettre à disposition des chercheurs, souvent après consultation de la communauté des utilisateurs comme à l'Ina ou à la BnF, des outils supplémentaires d'interprétation et d'interrogation des sites archivés.

Le plus attendu a probablement été la possibilité d'une interrogation en plein texte[6] des ressources archivées qui permet d'échapper à une consultation où domine la « *single URL approach* ». Les archives portugaises, françaises (Ina et BnF) ou encore britanniques et japonaises y ont recours.

Cette possibilité enrichit l'expérience de navigation à deux titres au moins. D'abord, la recherche en plein texte permet de thématiser des recherches qui n'auraient pu aboutir par une consultation « à la main » des sites, l'un après l'autre. C'est une étape fondamentale dans la constitution des corpus de recherche et l'émergence de nouveaux objets. Ensuite les résultats obtenus permettent des tris multiples (dates, occurrences d'un terme, d'une expression, présence d'un type de ressources, etc.).

Dernière étape en date de l'évolution des interfaces, la mise en place d'une multitude de « surcouches » de recherche et de manipulation des données qui permettent d'exploiter l'archive et d'en rendre compte sous une forme particulière. En Grande-Bretagne, le moteur Shine[7] permet par exemple de soumettre les résultats d'une recherche à un traitement statistique et de générer une représentation sous une forme proche de Google Ngram. L'archivage

6. La recherche en « plein texte » est ici employée pour traduire l'anglais « *full-text search* ».
7. https://www.webarchive.org.uk/shine.

des réseaux socionumériques par la BnF et l'Ina permet le traitement des métadonnées associées aux messages collectés et donne la possibilité d'interroger les données collectées de manière croisée (par exemple par mot-dièse et langue ou date, etc.) et de représenter les résultats de multiples façons : frises chronologiques, nuage de mots, liste d'emojis les plus utilisés, etc.

Enfin, des initiatives émergent en périphérie d'Internet Archive et des grandes institutions d'archivage pour donner accès à des outils permettant de nouvelles exploitations des données sauvegardées. Citons par exemple Internet Archive Wayback Machine Link Ripper[8] qui permet de retrouver toutes les URLs archivées dans Internet Archive à partir d'une URL connue ; WebART (pour Web Archives Retrieval Tools) qui est un ensemble d'outils et d'interfaces de recherche proposé par l'équipe Dutch Web Archive de la bibliothèque nationale des Pays-Bas et le Centrum voor Wiskunde en Informatica de l'université d'Amsterdam[9], parmi lesquels on trouve WebARtist, un moteur de recherche en plein texte capable de prendre en compte les temporalités pour retrouver un texte ou une image ; ou encore Wayfinder de Megan Dougherty qui permet de personnaliser son interface de recherche dans les archives du Web en complément de la suite WebArchivist (Dougherty, 2017).

Des outils d'analyse

Si la recherche par mots-clés peut sembler indispensable à des chercheurs, habitués, comme le grand public, aux moteurs de recherche et au plein texte, la fourniture de ces fonctionnalités n'est pourtant pas une évidence. C'est seulement en 2016 que la BnF va implémenter une recherche en plein texte dans ses archives du Web des années 1990, puis dans sa collecte des attentats de 2015, et permettre une recherche avancée par mots-clés, dates, auteurs ou types de formats (.html, .pdf,

8. https://tools.digitalmethods.net/beta/internetArchiveWaybackMachineLinkRipper.
9. http://www.webarchiving.nl/news.

etc.) en adaptant le moteur de recherche utilisé par la British Library, Shine. D'autres indexations sont en cours, mais une partie de la collection de la BnF reste interrogeable seulement en connaissant l'URL du site recherché. La Wayback Machine d'Internet Archive ne fournissait pas non plus de recherche autre qu'une recherche par URL jusqu'à une période récente. Sa recherche par mots-clés comporte par ailleurs le biais de ne fouiller que les pages d'accueil des sites archivés.

Deux remarques s'imposent :

- la première est qu'il faut composer avec des archives en constante évolution, tant par leur mode d'archivage que d'interrogation. Les outils et fonctionnalités offerts par les organisations évoluent au cours même d'un projet et peuvent rendre caduques des méthodologies ou les faire évoluer. Ainsi notre projet Web90 a commencé en 2014 sans autre possibilité de consultation des archives des années 1990 que la recherche par URL (à part pour celles conservées à l'Ina, qui avaient déjà une recherche en plein texte). Quand, en 2016, la recherche en plein texte devient possible, aux heures passées à chercher des sites susceptibles de fournir des informations sur un sujet précis succède une quasi-instantanéité d'accès à des résultats plus variés et détaillés – sans toutefois faire disparaître les biais documentaires, puisque ces résultats comportent des choix introduits dans la conception du moteur de recherche.
- la seconde est le souci des institutions de valoriser ce patrimoine nativement numérique, de le rendre exploitable en fournissant des outils de fouille. Plusieurs éléments expliquent ce choix. Comme le note Thomas Drugeon (DL Web Ina), le chercheur ne peut pas partir avec les données, les sortir des enceintes des bibliothèques en France. Les outils d'analyse se doivent donc aussi d'être disponibles dans l'enceinte de consultation, et ils sont parfois nécessaires pour permettre la lisibilité de plusieurs milliers d'éléments (sites, pages, *hashtags,* etc.) ou les mettre en relation (par exemple au travers d'une recherche linguistique). Si les archives d'Internet Archive

sont en ligne et si on peut avoir le sentiment de pouvoir utiliser plus d'outils ou de les choisir, l'accès aux fichiers WARC[10] n'est pas acquis, et des contraintes techniques (mais aussi économiques) peuvent se poser.

L'évocation des fichiers WARC renvoie à des pratiques de traitement de données et métadonnées standardisées par le moyen d'outils informatiques (logiciels d'analyse lexicographique par exemple) qui s'apparentent à ce que Franco Moretti a qualifié de *distant reading* (lecture distante), proposant :

> « *What we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them*[11]. » (Moretti, 2013)

Loin de constituer la seule forme de lecture possible des archives du Web, la lecture distante permet toutefois dans le cadre de grands corpus d'avoir un aperçu que les capacités humaines de lecture ne permettent pas.

Les archives du Web passent ainsi sous le « macroscope historien » (Graham *et al.*, 2015). Des outils d'analyse en accès ouvert comme Iramuteq ou Gephi, ou développés par les institutions (pour produire par exemple des *timelines*, des diagrammes représentant les emojis ou images les plus tweetés dans les archives de l'Ina) permettent d'entrer dans les masses documentaires, par le contenu textuel, mais aussi par les images, les émoticônes ou encore les *hashtags* pour Twitter.

La lecture distante a été notamment utilisée pour la reconstruction de Geocities par Ian Milligan (2012-2017). Il a par exemple extrait des images afin de mesurer les promiscuités et récurrences visuelles au sein de ce service de pages personnelles particulièrement populaires dans les années 1990.

10. Le format WARC (Web ARChive), largement adopté depuis le milieu des années 2010, en remplacement de son prédécesseur, le format ARC, permet d'établir des standards en matière de collecte et de stockage des données hétérogènes présentes sur Internet. Pour plus de précisions, voir : http://www.bnf.fr/fr/professionnels/dlweb_boite_outils/a.dlweb_formats_fichiers.html.

11. « Ce qu'il nous faut, c'est un petit pacte avec le diable : nous savons comment lire les textes, apprenons maintenant comment ne pas les lire. » (Notre traduction.)

Les approches inspirées des *cultural* et des *visual studies* d'Anat Ben-David (reconstruction de noms de domaine disparus tel le .yu de l'ex-Yougoslavie, ou analyse de la couleur des domaines nationaux, voir Ben-David 2016 ; Ben-David *et al.*) contribuent également à apporter un nouveau souffle (et de la couleur) dans un paysage académique qui reste par ailleurs toujours très marqué par des approches linguistiques ou politiques, ce que relevaient déjà Dougherty *et al.* il y a quelques années (2010).

Outre le développement d'outils au sein du monde de la recherche, qui doit permettre aux chercheurs d'accéder à de plus en plus de boîtes à outils (voir par exemple The Archives Unleashed Project[12]), les bibliothèques ont également développé des plateformes de consultation, que ce soit la British Library, la BnF, la Bibliothèque royale du Danemark ou l'Ina. Elles sont susceptibles de prendre en charge l'outillage de la recherche à toutes les phases de celle-ci, depuis la recherche dans les fonds (recherche avancée, sélection par facettes de dates, noms de domaine, etc.), puis l'analyse (chronologies, graphiques, statistiques, représentations de tendances linguistiques sur le modèle de Google Ngram) jusqu'à la préservation, voire le partage du corpus.

Élargissant la thématique au-delà des archives du Web, pour considérer les données numériques susceptibles d'être analysées au sein de la bibliothèque de manière plus générale, la BnF a ainsi lancé une enquête prospective en 2017 pour préfigurer un nouveau service de fourniture de données à destination de la recherche, appelé provisoirement Laboratoire d'étude et d'analyse de corpus numériques (Moiraghi, 2018). Un autre exemple récent de ces efforts est fourni par la réalisation à la Bibliothèque royale danoise d'une nouvelle interface (voir « A wayback machine for the UKWA Solr based warc-indexer framework[13] ») incluant, de la recherche à la visualisation des résultats, de multiples fonctionnalités, type cartographie interactive de liens ou encore localisation des images et temporalités des collectes.

12. https://archivesunleashed.org.
13. Pour un descriptif et des captures d'écran de l'interface du projet, voir https://github.com/netarchivesuite/solrwayback.

La situation du chercheur en 2018 face aux archives du Web n'a ainsi plus rien à voir avec celle du début de la décennie. Reste que ces outils, s'ils peuvent simplifier la recherche, impliquent aussi de penser les biais et la couche de médiation supplémentaire qu'ils introduisent. Les travaux de Noortje Marres (2012, 2015), de Bernhard Rieder et Theo Röhle (2012) notamment, ont montré que le chercheur en sciences sociales doit conserver une distance critique face aux « présupposés épistémologiques contenus dans les outils » (Mabi, Plantin et Monnoyer-Smith, 2014).

On rappellera à cet égard les réflexions stimulantes d'Anat Ben-David et Hugo Huuderman sur les moteurs de recherche dédiés aux archives du Web (2014), ou de Megan Ankerson (2015a) sur les interfaces de consultation des archives du Web. Les outils d'analyse donnent également matière à réflexions méthodologiques, par exemple dans les travaux liés à la reconstruction de Geocities (Milligan, 2017) ou de domaines nationaux (Brügger, 2017a ; Brügger, Laursen et Nielsen, 2017).

Dans le panorama des différents outils d'exploitation des archives du Web, une situation particulière s'est présentée lors des collectes « d'urgence » qui ont suivi les attentats parisiens autour de *Charlie Hebdo* et ceux du 13 novembre 2015 : elle a amené la BnF et l'Ina à questionner leurs outils. En effet, si d'importants moyens techniques et humains ont été mis en œuvre lors de la collecte, la nécessité d'outils d'analyse performants s'est posée clairement face à ces collectes de grande ampleur.

La BnF a ainsi fait le choix de tester l'implémentation de la recherche en plein texte dans son corpus ; l'Ina a, de son côté, travaillé à fournir des outils, notamment de visualisation, pour exploiter les données et métadonnées du sien[14]. Les entretiens menés avec les porteurs de ces initiatives institutionnelles révèlent que celles-ci se trouvent souvent face à une tension :

> « dans la majorité des cas, les usagers qui viennent consulter un fonds du dépôt légal du Web le considèrent comme un fonds parmi d'autres au sein de

14. Parmi les fonctionnalités proposées : la possibilité de croiser plusieurs éléments tels des mots-dièses, mots-clés, statistiques de langues ou encore nombre de retweets.

> leurs recherches, ils ne vont pas dépenser une énergie énorme pour comprendre les limites. Mais certains vont chercher à aller plus loin. Nous sommes tiraillés entre ces besoins pointus et ceux de la majorité des usagers, pour lesquels il ne faut pas trop spécialiser l'outil, sinon il devient incompréhensible[15]. »

À n'en pas douter, en fournissant à la fois les données et les outils pour les exploiter, les institutions d'archivage assument un rôle central. Le chercheur se doit donc de déployer une vigilance et un effort pour comprendre à la fois les apports et biais des corpus, mais aussi ceux des outils fournis, en gardant à l'esprit que la neutralité des données comme celle des outils est illusoire (Plantin et Monnoyer-Smith, 2013). Dans le même temps, la mise en place de projets de recherche qui permettent aux chercheurs de signaler des URLs à archiver au moyen de l'outil BnF Collecte du Web, ou des ateliers du DL Web Ina, montre une attention aux besoins des chercheurs et aux contributions qu'ils peuvent apporter dans le cadre de l'exploitation du patrimoine numérique ; les institutions cherchent à penser leurs publics et saisir leurs demandes parfois très différentes.

Penser l'archive du Web en contexte

Figure 1- Mème circulant largement sur la Toile

15. Entretien avec Thomas Drugeon (responsable du DL Web à l'Ina), mené par V. Schafer et M. Borelli le 21 mars 2016 (https://asap.hypotheses.org/tag/ina).

Un mème[16] valant parfois mieux qu'un long discours, celui qui illustre ce début de section rappelle combien la prise en compte du contexte se révèle indispensable pour prétendre à une réelle compréhension de l'archive numérique.

Bien que vrai en soi puisque ce mème reproduit le résultat d'une recherche réellement effectuée et devenue virale, son propos ne l'est qu'à l'aune du rapprochement des différentes requêtes des utilisateurs fait par le moteur de recherche de Google.

L'archive du Web est issue d'un contexte global de production. La grande simplicité de la structure des sites des premières années du Web (des années 1990 au début des années 2000 dans la majorité des pays occidentaux) nous rappelle par exemple de quel poids pesaient encore les offres d'abonnement à la minute sur la consultation et par conséquent sur l'offre informationnelle proposée à l'internaute. Associée aux débits offerts par les modems de l'époque, cette structure des coûts de consultation explique en partie la faible profondeur des sites et la place marginale des images qui ne peuvent en conséquence être analysées hors de ces contraintes externes au Web lui-même.

Dans le contexte actuel, la production des contenus « générés par les utilisateurs » (*user-generated content*) est aussi influencée, dans une large mesure, par les dispositifs eux-mêmes qui récoltent, traitent et analysent ces données, invitant en outre à aimer, retweeter, etc. L'activité « dynamique » et automatique de nombre d'outils web, notamment de robots, doit également être prise en compte pour cerner la complexité du Web contemporain.

Si l'on pousse plus loin la prise en compte des agents techniques, ce qui apparaît à la surface de la page n'est que le rendu visuel d'un ensemble de codes informatiques. Ces derniers, à commencer par le .html, contiennent non seulement la trace des opérations de formatage des données et des logiciels, mais aussi des informations qui peuvent aller au-delà des paramètres techniques et relèvent des contextes de production.

16. Un mème internet est un élément de contenu (sous la forme de texte, image fixe ou animée, ou encore son, et selon des formats très divers) repris et décliné massivement sur la Toile (parfois transformé d'un format à un autre).

Il faut également faire une place aux contextes de réception des sources archivées. L'analyse quantitative de Twitter nous en donne un exemple saisissant. Comment juger de l'importance d'un tweet ou d'une série de tweets ? Faut-il l'analyser à l'aune de sa place dans l'espace de communication du réseau (ses retweets, ses likes, etc.) ? Selon quelle métrique ? Faut-il éventuellement s'ouvrir à une dimension plurimédiatique en soulignant que certains messages, du fait de la notoriété de leur auteur ou de son ancrage dans une communauté spécifique, connaissent un écho important hors du réseau lui-même (on pense en particulier aux relais que les journalistes offrent à certains messages dans un article, un journal télévisé, une émission de radio dont les printemps arabes ont été un exemple poussé jusqu'à l'absurde[17]) ? Impossible de décontextualiser totalement l'analyse pour faire du tweet un élément parmi d'autres. Bien entendu, la lecture distante propose une autre approche des corpus en faisant émerger des relations entre entités et groupes. Mais elle ne peut faire l'économie de la lecture attentive (close reading), sous peine de décontextualisation. Rendre compte d'un contexte global, ce n'est pas se tenir à distance, c'est rendre compte d'un va-et-vient entre les échelles de lecture et de compréhension d'un corpus.

Le contexte de réception est aussi fortement influencé par la structure en réseau du Web et de ses archives. La viralité des informations, leur reprise et leur modification entre sites et même entre pages est un élément d'appréciation de contexte important, comme le souligne Clément Oury (2012). Leur instabilité et leur volatilité en sont un autre, non négligeable. Une consultation des archives gagne à inclure une réflexion sur ce qui n'est pas archivé, ou ce qui risque de ne pas l'être. Une page web avec une série de liens vers des documents non archivés travaille la suggestion, l'évocation – voire la frustration du lecteur. Le chercheur doit travailler « en creux », multiplier les

17. Quand l'Occident relaie les contestations qui émergent à partir de fin 2010 en Tunisie, le rôle d'Internet et des réseaux sociaux fait l'objet d'analyses enthousiastes qui relèvent souvent du *solutionnisme* technologique (Morozov 2014), c'est-à-dire d'une pensée qui prête aux nouvelles technologies la capacité à résoudre tous les grands problèmes, de la faim dans le monde à la maladie. Les espoirs sont rapidement déçus questionnant l'impact réel des mobilisations en ligne (Bortzmeyer, 2016).

sources et ne pas s'en tenir à l'illusion d'une archive universelle et exhaustive. Au-delà des archives web, la presse spécialisée, des entretiens oraux ou les archives audiovisuelles livrent ainsi de multiples pistes pour reconstituer l'histoire du Web (Schafer, 2015).

Enfin, le contexte d'archivage informe sur le traitement donné à l'archive du Web. Une collecte n'est jamais une sauvegarde neutre des données : c'est une construction d'événements préjugés. Lorsqu'une collecte est décidée pour documenter un événement, une période ou un sujet d'intérêt, un ensemble de critères est mis en place pour sélectionner ce qui sera conservé. Comme l'archiviste le fait avec les masses de papier qui lui parviennent, sans en prendre exhaustivement connaissance, un tri est effectué a priori. Ainsi, il a été choisi par les institutions françaises d'archivage de poursuivre un objectif de représentativité et non d'exhaustivité en matière de conservation des sites des partis durant les campagnes électorales : les sites des petits partis aux extrémités du spectre politique sont conservés pour que l'ensemble soit représentatif des équilibres du spectre et non du poids respectif des formations en ligne.

Outre-Atlantique, d'autres considérations commencent à entrer en ligne de compte. Notamment sur le plan politique, certains acteurs entendent créer dans les corpus conservés une dimension « non oppressive », c'est-à-dire faire une place clairement identifiée et assumée à des groupes et des individus minoritaires au sein de la société et des flux de données en ligne. Le projet Documenting the Now[18] organise ainsi depuis 2016 une collecte des archives de Twitter selon des thématiques choisies en matière de genre, de critères « ethno-raciaux » anglo-saxons (dans le cadre entre autres du mouvement Black Lives Matter) ou de diversité culturelle.

Cette question des équilibres et de la représentativité est bien entendu critique pour les institutions en charge de la conservation, et pose des questions de fracture numérique, comme on a pu le montrer précédemment. Les faiblesses de la représentation en ligne des Suds (Gomes *et al.*, 2011) préoccupent l'IIPC. Les archivages des domaines .ao et .cv (angolais et cap-verdien)

18. https://www.docnow.io.

par les institutions portugaises en vertu de l'histoire coloniale du pays questionnent quant à eux les logiques éventuelles d'appropriation culturelle que risquent de faire émerger ces pratiques.

Si cette question de la bonne pratique en matière de construction des collections n'est pas tranchée et ne le sera probablement jamais de manière totalement satisfaisante, la participation des usagers des archives semble constituer une voie féconde d'amélioration. Cette association des chercheurs et des usagers au processus de collecte et aux règles qui la gouvernent se multiplie, à l'image des pratiques de la BnF et de l'Ina. Bien entendu, cette inclusion n'est pas nouvelle : de Michelet, chef de la section historique aux Archives nationales, à Jean-Noël Jeannenet, président de la Bibliothèque nationale de France, l'historien en particulier a toujours eu à cœur de participer aux politiques de conservation de son temps par la coconstruction des contextes d'archivage.

Les divers enjeux posés par les contextes de production, de réception et d'archivage illustrent la multiplicité des problématiques qu'il s'agit d'entrelacer au cœur des analyses qui prennent l'archive du Web comme support. Loin de bouleverser les règles traditionnelles de l'analyse, le contexte continue d'enrichir la compréhension du contemporain et, demain, d'un passé dont les traces sont d'ores et déjà lisibles en ligne.

UNE RECHERCHE AUX INTERFACES

Les éléments qui précèdent ont permis de montrer l'extrême variété des archives du Web, l'intrication d'enjeux techniques, politiques, sociaux et économiques qui influent sur les fonds constitués et mis à disposition. À la variété des archives répond la variété des méthodologies et approches, aussi vaste que le champ des questionnements qui peuvent prendre l'archive du Web pour objet.

Les archives du Web, quels publics ?

Fin 2010-début 2011 la délégation à la Stratégie et à la Recherche de la BnF lançait une enquête qualitative auprès de publics potentiels des archives du Web (Chevallier et Illien, 2011). Elle identifiait alors trois profils : les chercheurs (en histoire, philosophie et sociologie des sciences et des techniques notamment), ensuite les professionnels (avocat, consultant marketing, documentaliste, ingénieur brevet, journaliste), enfin le tout-venant de la bibliothèque de recherche.

Une enquête en ligne menée notamment auprès des chercheurs autour des ateliers du DL Web Ina montrait quant à elle en 2011 :

> « […] encore une certaine défiance [de leur part] autour des critères de fiabilité, d'autorité et d'instabilité. Les rares pratiques d'archivage étaient en prise directe avec le Web vivant (*bookmarks, screencast*[1]…) ou bricolées d'après un modèle imprimé (.pdf), sans stratégie archivistique ou documentaire. » (Merzeau et Mussou, 2017)

1. Les *bookmarks* sont des systèmes de marque-pages ou favoris permettant de retrouver les pages ou sites jugés intéressants. Quant au *screencast*, il désigne l'enregistrement vidéo numérique de l'affichage d'un écran.

L'étude de Meghan Dougherty *et al.* (2010) avait également fourni une base solide de réflexion. Alors que ses auteurs notaient le fossé entre la potentielle communauté de recherche et sa réalité, bien plus modeste, ils proposaient pour l'élargir une série de recommandations qui restent d'actualité, que ce soit du tutorat et des formations, des appels à projets, le développement et la mise à disposition d'outils d'analyse, etc.

Si Chevallier et Illien (2011) notaient surtout des besoins ponctuels en 2010-2011, les quelques années qui nous séparent de cette enquête prospective ont permis de voir un intérêt croissant pour les archives du Web, dans la recherche comme dans les médias, même si cet intérêt ne se lit pas toujours en termes de consultation dans les enceintes de la bibliothèque. Et le constat dépasse bien sûr les frontières hexagonales. Comme le relevait en effet l'historienne britannique Jane Winters (2017a) :

> « *Anyone who works with web archives quickly becomes used to the fact that most people have not even heard of them – even fewer understand what they are and where you might be able to access them. In 2016, however, it seemed as though web archives began to filter into the public consciousness, to move from the technology pages of the more serious newspapers to the political and even cultural sections*[2]. »

L'année 2016 aura-t-elle été celle des archives du Web ? En France, comme dans le monde anglo-saxon, ce sujet jusque-là plutôt confidentiel aura en tout cas fait l'objet d'une large couverture médiatique, notamment de la part du *Monde*, de *Libération* ou encore de *L'Express*, à la faveur des vingt ans de la fondation étatsunienne Internet Archive et des dix ans du dépôt légal du Web en France[3].

2. « Toute personne travaillant avec des archives web s'habitue rapidement au fait que la plupart des gens n'en ont même pas entendu parler – encore moins comprennent ce qu'elles sont et où y accéder. En 2016, cependant, il semble que les archives web commencent à filtrer dans la conscience publique, à passer des pages technologiques des journaux les plus sérieux aux sections politiques et même culturelles. » (Notre traduction.)
3. http://bnf.hypotheses.org/1105.

Cependant une analyse plus fine des publics intéressés, ou même de la croissance de l'audience, reste complexe. Ainsi, alors que la croissance des consultations générales d'archive.org montre une nette évolution depuis 2009[4], les archives du Web ne sont prises en compte qu'à partir de 2013 dans ces données et les statistiques mises à disposition ne sont pas toujours aisément interprétables.

Reste que du côté des bibliothèques, la marge de progression de la fréquentation est encore réelle et à attendre dans les prochaines années, alors que les premières thèses d'histoire utilisant des archives du Web institutionnelles, à l'instar de celle de Sophie Gebeil (2015) sur les mémoires de l'immigration maghrébine sur le Web, ont ouvert la voie. Le développement de l'accès aux archives du Web en région par la BnF et l'Ina garantit la possibilité d'accéder à ces archives sur tout le territoire national.

Si un travail de pédagogie s'impose dans le monde académique et l'enseignement supérieur pour former les étudiants et les inciter à consulter ces archives, les institutions d'archivage ont bien compris qu'elles pouvaient stimuler l'intérêt par des appels à chercheurs, comme le font la BnF, l'Ina ou encore la British Library.

Ainsi le projet The Big UK Domain Data for the Arts and Humanities project (BUDDAH[5]) a-t-il recruté dix jeunes chercheurs issus des humanités en 2014 pour leur proposer de développer au sein de la British Library des projets de recherche fondés sur les archives du Web. En sont ressortis des travaux féconds (Winters, 2017), à l'instar de ceux de Marta Musso qui a analysé les premiers pas des sites web britanniques sur la Toile (Musso et Merletti, 2016).

D'un point de vue plus qualitatif, les usages et besoins de recherche commencent à être de mieux en mieux cernés, que ce soit par des initiatives comme les ateliers du DL Web Ina menés par Louise Merzeau et Claude Mussou pendant

4. https://blog.archive.org/2015/01/26/archive-org-download-counts-of-collections-of-items-updates-and-fixes/.
5. https://buddah.projects.history.ac.uk.

six ans[6], des rapports comme celui réalisé par le Net Lab danois de l'université d'Aarhus (Costea, 2018) ou l'étude prospective conduite par la BnF en 2017 (Moiraghi, 2018), même si les besoins sont loin d'être figés et homogènes.

Enfin il convient, au-delà des publics espérés et souhaités des archives du Web, de garder à l'esprit qu'il y a aussi des publics exclus des archives du Web. Certains gouvernements ont pu couper ponctuellement l'accès à Internet Archive : la Chine en 2014, le gouvernement russe en juin 2015 et la Jordanie en 2017 (Butler, 2017).

L'utilité des archives du Web dépasse les seules communautés de recherche et ces archives peuvent également susciter l'intérêt citoyen. Les journalistes, les juristes, mais aussi la société civile pourraient s'en emparer. Certains ont déjà commencé à le faire, comme ceux qui reprennent les tweets fondés sur les archives du Web d'Internet Archive qui scandent la présidence de Donald Trump depuis 2017 aux États-Unis, confrontant sa politique à ses annonces passées et documentant ses contradictions.

Les archives du Web : *trading zone* et objet interdisciplinaire

Les archives du Web sont le résultat de mobilisations hybrides d'innovateurs, d'utilisateurs et d'entrepreneurs, ainsi que d'une variété d'experts qui vont des chercheurs aux bibliothécaires et archivistes en passant par les informaticiens – chacun avec ses outils, attentes et cultures. Elles sont donc à plein titre un objet de recherche multi et interdisciplinaire, qui peut bénéficier des perspectives de disciplines telles que l'histoire, les sciences de l'information et de la communication, la sociologie, les sciences du langage, et bien sûr l'informatique, ainsi que de domaines disciplinaires comme les études sociales des sciences et des techniques (en anglais, *Science and Technology Studies* ou STS), ou encore les *media studies* et la linguistique de corpus. Si ces

6. Voir le bilan qu'elles en tiraient en 2017 : https://webcorpora.hypotheses.org/302.

regards portés sur les archives du Web peuvent varier (lecture davantage diachronique ou sémiotique, ou encore inspirée par les *visual studies*, etc.), les chercheurs empruntent de plus en plus de clés de lecture et d'analyse à d'autres champs, voire se retrouvent dans des tendances dépassant les disciplines, à l'instar des *code studies* ou des *Internet studies*. Des approches dérivées des sciences juridiques et politiques peuvent également être mobilisées utilement pour explorer les institutions de standardisation de l'archivage du Web et leurs alliances, ainsi que les questions de droit d'auteur et de régimes de propriété intellectuelle applicables aux contenus archivés (voir par exemple Dulong de Rosnay et Guadamuz, 2017).

Les approches relevant des STS, en dialogue avec d'autres domaines disciplinaires, sont par exemple fructueuses pour appréhender et analyser des aspects qui relèvent de la « boîte noire » des archives du Web. Cette approche permet ainsi de penser les « relations de pouvoir » au sens large qui sont inscrites dans les archives du Web (Badouard *et al.*, 2016), du rôle des GAFAM dans la constitution d'archives privées aux missions publiques d'institutions patrimoniales, en passant par la place des usagers dans les politiques conduites. Les notions de médiation, d'intermédiation, ainsi que celle d'*agency* (ou puissance d'agir[7]) peuvent être utilement appliquées au Web et à ses archives, pour observer des agencements au sein des dispositifs qui reflètent ces relations de pouvoir. Comme on l'a vu tout au long de cet ouvrage, les négociations humaines et techniques, à la fois au niveau de la collecte et de l'exploitation de l'archive du Web, incluent plusieurs opérations : des choix de fréquence de collecte, de périmètre et de profondeur ; des modalités de programmation des robots et des processus de dédoublage des données ; l'exclusion d'éléments spécifiques, comme les publicités ; ou encore la création de plateformes et d'environnements de consultation proposant chacun des designs et fonctionnalités différents. L'archivage du Web est le résultat du coformatage mutuel des contenus et des artefacts, des développeurs et des utilisateurs ; il découle d'un ensemble de pratiques et discours souvent triviaux et considérés comme acquis, qui jouent

7. Traduction proposée par Proulx, 2009.

pourtant un rôle dans la conception, la régulation et l'entretien du Web. Tous ces éléments se doivent d'être analysés comme le résultat des motivations, des choix, des alliances des acteurs de l'archivage du Web – et, en même temps, on ne peut pas se priver d'une analyse fine des aspects techniques et parfois économiques qui les sous-tendent (Schafer *et al.*, 2016).

Les archives du Web et le patrimoine numérique peuvent également être explorés à l'aune de la notion d'« objet-frontière », concept proposé par Susan Leigh Star et James Griesemer (1989) pour décrire ces processus où des acteurs provenant de différents milieux sociaux et politiques, et appelés à coopérer, arrivent à se coordonner malgré des points de vue divergents, établissant une compréhension mutuelle sans pour autant perdre de la diversité et de la richesse des origines (Trompette et Vinck, 2009, p. 6-7).

À cet égard, l'archivage du Web peut être considéré comme une zone d'échange (*trading zone*) au sens de Galison (1997). Cette métaphore a été utilisée à l'origine par son créateur afin de rendre compte de la manière dont des physiciens issus de différentes écoles arrivaient à collaborer entre eux et avec des ingénieurs pour développer des objets techniques complexes tels que le radar ou le détecteur de particules. Appliquée aux archives du Web, elle révèle aussi toute sa pertinence : ces archives sont des objets complexes au croisement non seulement de plusieurs disciplines, mais de plusieurs figures et communautés professionnelles (bibliothécaires, archivistes, ingénieurs et chercheurs).

Tendances de la recherche

Les recherches sur les archives du Web sont en plein essor, ce dont témoignent notamment la constitution du groupe de réflexion RESAW[8] à l'échelle européenne dans la première moitié de la décennie 2010, ou encore l'attractivité de

8. http://resaw.eu.
On notera avant des publications pionnières issues du monde de l'archivage du Web, à l'instar de celle coordonnée par Julien Masanès (2006).

manifestations type hackathons, dédiées aux archives du Web et organisées plus récemment des deux côtés de l'Atlantique.

Les travaux de recherche ont d'abord été dominés par des réflexions méthodologiques, initiées notamment par l'historien danois Niels Brügger. Ils ont commencé par souligner les défis que représente ce type de matériaux, insistant sur les médiations et reconstructions que subissent les archives du Web (Brügger, 2012b). Ces recherches ont par ailleurs intégré des enjeux disciplinaires et interdisciplinaires au fil des années, comme l'illustrent les réflexions menées sur les liens entre archives du Web, *Digital Studies* et *Digital Humanities* (Brügger, 2016). Par ailleurs, sous l'impulsion de courants de réflexion notamment liés aux *Science and Technology Studies*, des efforts ont également été faits pour comprendre la fabrique des archives du Web, comme nous venons de l'évoquer (Schafer, Musiani et Borelli, 2016).

Des enjeux de gouvernance aux enjeux de pouvoir… la frontière est évidemment ténue et les cas de censure ou encore de suppression d'archives comme celles menées en Grande-Bretagne par les conservateurs en 2013 (Winters, 2017b) ont également sensibilisé aux enjeux politiques et géopolitiques que posent ces archives (Schafer, 2017). Initialement très tournées vers les dimensions et usages internes à la recherche (Dougherty *et al.*, 2010), ces réflexions ont dans un second temps également été articulées avec des enjeux pédagogiques, qui montrent la volonté d'ouvrir ces archives à des publics plus larges (Winters, 2017).

Bien sûr les archives du Web ont aussi été insérées dans des réflexions épistémologiques et méthodologiques plus générales sur le patrimoine numérique (Treleani, 2017 ; Bachimont, 2017b). Elles bénéficient par ailleurs de réflexions dédiées à d'autres types de sources nativement numériques, tels les forums de discussion en ligne (Paloque-Bergès, 2018), et de la volonté de penser les silences des archives, qui négligent les publicités en ligne ou encore les spams, comme l'a montré Finn Brunton (2017).

Des recherches ont essayé d'évaluer les données manquantes au regard du Web vivant tel qu'il existait (Huuderman *et al.,* 2015 ; Hale, Blank et Alexander, 2017). Différents, mais non moins complémentaires des précédents, des efforts ont

également porté sur la possibilité de reconstituer des sites spécifiques (voir Nanni, 2017 pour le site de l'université de Bologne), voire des noms de domaine disparus, à l'instar du .yu de l'ex-Yougoslavie, exploré par Anat Ben-David (2016). Les archives du Web ont également servi d'appui à des travaux visant à retracer l'évolution d'un domaine national spécifique (nous pensons ici plus particulièrement aux travaux menés au Danemark, voir Brügger, 2017).

D'autres chercheurs, sans négliger ces dimensions méthodologiques, sont entrés au cœur des archives pour les exploiter au service de sujets de recherche, que ce soit pour étudier les cultures numériques et l'histoire du Web lui-même[9] ou encore pour aborder des sujets généraux mais représentés en ligne. C'est le cas par exemple des recherches de Sophie Gebeil dédiées aux mémoires de l'immigration maghrébine en ligne (Gebeil, 2017), des travaux que consacre Peter Webster (2018) à l'histoire des religions en exploitant les archives du Web, et de ceux de Richard Deswarte consacrés à l'euroscepticisme britannique[10] tel qu'il a pu se manifester sur la Toile[11]…

En France, on notera parmi les travaux précurseurs, à des fins historiques ou non, ceux menés autour de Dana Diminescu au sein du projet e-diasporas (Diminescu et Loveluck, 2014) ou encore par Valérie Beaudouin sur les commémorations de la Grande Guerre. Les mémoires en ligne et les commémorations semblent aujourd'hui un terrain de recherche où les usages des archives du Web sont pleinement assumés, ce dont témoignent également les travaux d'Enrico Natale (2017) ou de Frédéric Clavert (2018b) sur les commémorations de la Grande Guerre.

Dana Diminescu et son équipe, comme Valérie Beaudouin qui a travaillé dans le cadre d'un projet soutenu par la BnF

9. Voir notre projet ANR Web90 mené de 2014 à 2018, les travaux de Marta Musso et Franco Merletti (2016) sur l'arrivée des sites d'entreprises et commerciaux britanniques en ligne, ou les travaux de Ian Milligan, 2017 sur Geocities.
10. http://sas-space.sas.ac.uk/6103/.
11. Les travaux mentionnés ici ont la spécificité d'utiliser des archives du Web institutionnelles. Il faut évidemment aussi mentionner les nombreuses et précoces analyses du Web dans le champ des sciences de l'information et de la communication (voir Barats, 2013). Les chercheurs ont alors pour les besoins de leur recherche souvent réalisé leur propre conservation des pages web étudiées, notamment au moyen de captures d'écran. Ces archives créées par le chercheur lui-même sont intéressantes mais sortent du périmètre considéré ici.

(Beaudoin et Pehlivan, 2017), avaient précocement assumé une approche orientée vers ce que l'on ne qualifiait pas encore pleinement de *Digital Humanities*, mais qui déjà tirait parti des possibilités d'utiliser des outils d'exploration et de cartographie de la Toile, travaillant également à les inspirer et les enrichir (on pense ici à Gephi[12] par exemple). En parallèle, d'autres chercheurs ont privilégié des approches plus « micro », à l'instar de Sophie Gebeil ou de nos travaux au sein de l'équipe Web90 (Schafer, 2018) invitant à penser autant « the Historian's Macroscope » (Graham *et al.*, 2015) que le microscope, et rappelant que :

> « […] *that simply because collections of digital material are in many cases big data, which opens the possibility of asking and answering new types of research questions, this does not necessarily mean that they have to be approached as Big Data*[13]. » (Brügger, 2015, p. 11)

Quels enjeux éthiques et déontologiques ?

Si les archives du Web soulèvent des enjeux en termes de recherche, elles ne sont pas sans poser également des questions éthiques et déontologiques, que ce soit aux archivistes ou aux chercheurs. Parfois trop rapidement assimilées à un débat entre droit à la mémoire et droit à l'oubli, ces questions renvoient en fait à une réalité plus complexe (Dulong de Rosnay et Guadamuz, 2017 ; Jones, 2016). Des enjeux politico-éthiques se retrouvent ainsi à tous les stades du cycle de gestion, depuis le choix de ce qui est préservé jusqu'à l'exploitation des données, en passant par leurs conditions d'accès (Pabón Cadavid *et al.*, 2013).

12. Gephi naît dans le cadre du projet e-diaspora porté par Dana Diminescu. C'est un outil de visualisation de réseaux maintenant largement utilisé dans la communauté scientifique.
13. « [...] le simple fait que ces collections numériques sont souvent des masses de données, qui ouvrent la possibilité de leur poser ou de répondre à de nouvelles questions, n'implique pas nécessairement pour autant de les approcher comme des masses de données. » (Notre traduction.)

Luciana Duranti a rappelé, lors de la conférence « The Memory of the World in the Digital Age » (Unesco, 2012), les interrogations éthiques posées par l'émergence du numérique face aux cadres légaux existants. L'affaire WikiLeaks, par exemple, révèle une ambivalence des attentes en termes de droit (« *conflicting rights in the digital environment* ») alors qu'étaient mises à jour des données diplomatiques et militaires. Leur mise à disposition publique à travers des archives sauvages pointe très concrètement du doigt la nécessité de repenser des problématiques légales par l'éthique et vice versa : le rapport au secret, à la raison d'État et à la transparence, la confiance dans les données et les documents, leur traçabilité, et leur sécurisation, la délimitation et la redéfinition du domaine public, la standardisation et la gestion des droits d'auteur et des personnes…

Ces préoccupations se sont manifestées lors de journées d'étude et de conférences récentes, par exemple lors des initiatives de la Bibliothèque du Congrès en 2016[14] et du National Forum on Ethics and Archiving the Web[15] de mars 2018.

Elles accompagnent un mouvement de réflexion plus large sur l'utilisation des données, des blogs, des forums, des sites. Aussi peuvent-elles s'appuyer sur toute une production, notamment dans le champ des sciences de l'information et de la communication, sur le statut public ou privé des échanges en ligne, leur publicisation et mise en visibilité, les enjeux d'anonymisation ou encore de consentement, notamment étudiés par Guillaume Latzko-Toth et Madeleine Pastinelli (2013) ou Christine Thoër, Florence Millerand *et al.* (2012). Si, comme le notent Madeleine Pastinelli et Guillaume Latzko-Toth, « la frontière naguère intuitive entre ce qui relève de la vie privée et de l'expression dans la sphère publique est mise à mal par les nouvelles formes d'interaction médiatisée par ordinateur », les traces d'activité recueillies sur Internet étant de nature publique dans le cas de l'archivage du Web, il convient de dépasser cette dichotomie. Ce qui « ne veut pas dire pour autant que le chercheur soit exonéré de la responsabilité de veiller au bien-être des personnes qui sont l'objet de la recherche et, surtout, de se soucier d'éviter de leur

14. Voir http://www.loc.gov/loc/kluge/news/save-web-2016.html.
15. http://rhizome.org/editorial/2017/oct/24/open-call-national-forum-on-ethics-and-archiving-the-web/.

nuire. Et sur ce plan, la question qui se pose n'est ni celle de l'accessibilité des informations, ni celle des attentes des acteurs, mais bien plutôt celle du degré de publicité des informations et de l'effet qu'est susceptible d'avoir l'intervention du chercheur sur cette publicité » (Latzko-Toth et Pastinelli, 2013).

La tendance générale se fonde, depuis le début des années 2010 (Latzko-Toth et Proulx in Barats, 2013), d'une part sur l'attention accrue à la manière dont les utilisateurs du Web perçoivent leur propre production de contenus et de traces sur les réseaux (contre l'idée d'un statut « pseudo-objectif de leurs écrits ») ; d'autre part sur le respect d'une « intégrité contextuelle », c'est-à-dire la prise en compte du contexte de la production (par opposition au fait de tenter de deviner les intentions des producteurs). Ces postures ont été introduites et travaillées de manière pionnière par la communauté des *Internet studies*, qui a rassemblé depuis 2012 des préconisations d'utilisation éthique des données, documents et matériaux issus du Web dans des chartes et des guides sous l'égide de l'Association of Internet Researchers (AOIR)[16]. Les préconisations en faveur d'une éthique de l'utilisation des données ont à voir aussi bien avec la promotion d'une déontologie du chercheur (et des missions patrimoniales de long terme des institutions) qu'avec la critique de la récupération de ces données à visée de profits par les entreprises privées, réduisant les données à des marchandises (« *data as commodity* »).

Pour le chercheur comme pour l'archiviste, la question de garder ou détruire, mettre en évidence ou cacher certaines informations relève d'un choix éthique – qui peut se traduire par des choix légaux dans les institutions juridiques. Ainsi, le droit californien oblige depuis 2013 les fournisseurs d'accès à mettre à disposition des mineurs une « gomme numérique » (« *digital eraser*[17] »). Le droit peut aussi s'opposer aux conceptions plurielles de la valeur de l'archive. Ainsi, l'association des

16. Par exemple, « AoIR: Ethical Decision-Making and Internet Research de 2012 » : https://aoir.org/reports/ethics2.pdf. Voir plus généralement https://aoir.org/ethics/.

17. Voir l'article de 2013 dans *TechCrunch* de Gregory Fereinstein sur cette possibilité offerte aux jeunes de demander à leur FAI la suppression de certains contenus: https://techcrunch.com/2013/09/24/on-californias-bizarre-internet-eraser-law-for-teenagers/.

archivistes français a dû mener bataille contre la Commission européenne qui proposait une loi pour la destruction systématique des données personnelles dans les archives numériques (affaire #EUdataP, depuis 2013).

Ian Milligan, traitant des millions de pages archivées de Geocities qu'il étudie, se pose dès lors la question : « *How can we ethically navigate the records of seven million people*[18] *?* » Outre qu'il suggère la lecture distante, pour ne pas centrer l'attention sur l'individu mais davantage sur la somme de ceux-ci, il propose également d'essayer d'évaluer le degré d'attente des acteurs face à leurs données personnelles, leur « *expectation of privacy* ». Alors qu'en termes d'éthique l'attention est souvent portée sur la protection de la vie privée, des données personnelles, il évoque implicitement le droit à la mémoire face au droit à l'oubli. Il souligne en effet :

> « *Leaving people out isn't ethical either.*
>
> *I feel similarly uncomfortable with leaving the voices of everyday people completely outside the historical record when there is ample opportunity to include them. Moving to a full opt-in process would likely lead to the historical record being dominated by corporations, celebrities and other powerful people, tech males, and those wanted their public face and history to be seen a particular way*[19]*.* » (Milligan, 2018)

Ajoutons dans le cadre des archives du Web, mais aussi des *newsgroups* (forums de discussion[20]) des années 1980-1990, un écart temporel entre production et exploitation des données qui a des conséquences. Que cela concerne :

18. « La question qui se pose dès lors est comment pouvons-nous étudier de manière éthique les enregistrements de 7 millions de personnes ? » (Notre traduction.)
19. « Laisser les gens à l'écart n'est pas éthique non plus. Je suis tout aussi gêné à l'idée de laisser les voix des gens ordinaires complètement à l'écart des documents historiques quand il y a amplement l'opportunité de les inclure. Le passage à un processus d'*opt-in* complet conduirait probablement à des dossiers historiques dominés par des sociétés, des célébrités et d'autres personnes de pouvoir, des mâles tournés vers les technologies, et ceux qui souhaitent que leur figure publique et leur histoire soient perçues d'une certaine façon. » (Notre traduction.)
20. Voir par exemple les travaux de Paloque-Bergès (2017, 2018).

- la possibilité de retrouver les personnes vingt ans après et d'obtenir un consentement ;
- des propos échangés dans le cadre d'une Toile ou de forums plus confidentiels alors, mais aussi en une période de tâtonnement sur les caractéristiques des échanges en réseaux ;
- des prises de position parfois très libres et provocatrices face aux premières velléités de mise en procès de FAI, d'hébergeurs, de censure. Certains des acteurs de ces débats ont poursuivi leur carrière dans le domaine du numérique et occupent aujourd'hui des positions institutionnelles et entrepreneuriales éloignées de leurs premières prises de position ;
- le peu de recours au pseudonymat, mal perçu dans les premiers échanges sur les *newsgroups*.

Ces éléments impliquent des précautions au stade de l'analyse et de la diffusion de la recherche. La question de la diffusion pose aussi celle de la transparence en matière de création ou de partage de corpus. S'il est ainsi possible de référencer les archives, il est difficile d'une part de les reproduire, de les partager (par exemple dans le cadre du dépôt légal français) et l'accessibilité des archives dans la Wayback Machine ne garantit pas non plus le droit de reproduction/réutilisation, en raison des questions de droits d'auteur (Milligan, 2016).

Défini par l'Unesco dans l'article 9 de sa charte sur la conservation du patrimoine numérique comme un patrimoine culturel devant être conservé et rendu accessible pour donner au fil du temps une image équilibrée et équitable de tous les peuples, nations et cultures (Schafer, Musiani et Borelli, 2017), le patrimoine numérique auquel appartiennent les archives du Web risque-t-il par ailleurs de reproduire une fracture numérique Nord/Sud (Gomes *et al.,* 2011) ? Comme nous l'avons précédemment évoqué, les membres de l'IIPC sont presque tous issus des pays les plus développés. Si les archives portugaises collectent, en vertu de l'héritage historique du pays, les .ao et .cv (Angola et Cap-Vert), certains pays ou régions du Monde (Inde, Afrique, Moyen-Orient) ne disposent pas encore d'archivage du Web

structuré. La situation évolue rapidement : le Chili, l'Afrique du Sud, la Chine ou encore la Malaisie ont pris des initiatives en ce sens par exemple. Mais le constat et les préconisations de Nicholas Taylor fin 2015 restent valables :

> « *The institutional membership of the IIPC, the comparatively high degree of professional activity in North America and Western Europe, and perhaps even the distribution of archival coverage in the Internet Archive Wayback Machine suggest that the opportunity gap may not just be in the volume of preserved content but also its diversity.* [...]
>
> *All of which is to highlight the need for community efforts to both expand the base and enhance the capacity of web archiving organizations, through a combination of interoperating local tools and/or third-party systems*[21]. » (Taylor, 2015)

Notons par ailleurs face aux cas de censure précédemment relevés (la Chine en 2014, le gouvernement russe en juin 2015, en 2016 la Jordanie) que ces exemples ne doivent pas nous sembler lointains : les velléités d'expurger le Web et ses archives peuvent aussi se manifester aux États-Unis ou en Europe. Jane Winters (2017b) rappelle ainsi l'initiative du parti conservateur britannique en 2013 de supprimer de son site dix ans de discours et de bloquer l'accès à Internet Archive.

D'autres enjeux émergent, par exemple lorsque les organisateurs du National Forum on Ethics and Archiving the Web soulignent en 2018 leur souhait de créer des archives plus riches, « non oppressives », davantage au service des publics et de l'histoire. Depuis au moins les années 2010, des mouvements de contestation sociale radicale comme les projets américains Occupy et Living Archives avaient en effet soulevé la question

21. « La composition institutionnelle de l'IIPC, le niveau relativement élevé d'activité professionnelle en Amérique du Nord et en Europe occidentale et peut-être même la distribution de la couverture archivistique dans la Wayback Machine d'Internet Archive suggèrent que le fossé ne se limite pas au volume de contenu préservé, mais aussi à sa diversité. [...] Tout cela pour souligner le besoin d'efforts communs pour élargir la base et améliorer la capacité des organisations d'archivage du Web, grâce à une combinaison d'outils locaux interopérables et de systèmes fournis par des tiers. » (Notre traduction.)

non seulement de la documentation, mais aussi de la place des populations minoritaires ou opprimées dans cette documentation. En effet, les archives numériques, « mémoire numérique sociale » (« *digital social memory* ») sont des outils sensibles et ambivalents des problématiques politiques : elles peuvent servir de preuves juridiques dans le cadre d'affaires légales, mais aussi devenir des porte-voix puissants de causes, des supports de mémoires de la communauté à des moments de changement, de crise, voire de révolution, comme au moment des printemps arabes. A contrario, elles pourraient aussi faciliter la surveillance et renforcer des logiques de traçabilité, voire d'oppression, en permettant de retrouver notamment des prises de position politiques, éthiques, religieuses, etc. Les enjeux de représentativité des archives sont ainsi posés avec acuité dans le projet Documenting the Now, lancé en 2016 et porté par plusieurs institutions universitaires des États-Unis. Visant à développer notamment une application ouverte permettant de préserver, collecter et analyser les contenus de Twitter, ses concepteurs revendiquent aussi le souci de lutter contre les silences des archives. La page d'accueil du projet s'ouvre sur des photographies liées aux manifestations et émeutes de Ferguson, suite à l'affaire Michael Brown, jeune Afro-américain abattu en août 2014 par un policier, et au mouvement Black Lives Matter. Des enjeux de genre et de diversité culturelle ne manqueront pas non plus de traverser les problématiques d'archivage.

CONCLUSION

En dépit du caractère récent de l'histoire de ses processus de conservation, l'archive du Web suscite au sein de la communauté scientifique un intérêt croissant. En effet, peu d'objets, de périmètres d'étude et de réflexions contemporaines peuvent aujourd'hui se tenir « hors champ », comme l'aurait dit Louise Merzeau, alors même qu'Internet est devenu le phénomène massif que l'on connaît. À sa mesure, cet ouvrage se propose d'être un jalon pour découvrir, réfléchir et s'approprier l'archive du Web.

D'abord, le patrimoine numérique et nativement numérique constitue un élément décisif de notre modernité occidentale. Indiscutablement, le Web joue en effet maintenant un rôle dans nos mémoires individuelles et collectives et sa conservation constitue un enjeu patrimonial de premier plan. Les institutions qui, en France comme en Europe, ont été chargées de cette tâche ont su, dans le sillage des pionniers américains, mais à leur manière, trouver les moyens de constituer les corpus mis à disposition du public et des chercheurs depuis quelques années déjà. Collecte difficile, qui s'affronte à une nature particulière de l'archive, remarquable par sa masse impressionnante et toujours en augmentation, par les lacunes qu'elle comporte (jusque sur les pages archivées elles-mêmes) et l'hypertextualité qui relie ses différents éléments constitutifs (pages, ressources, etc.). À ces caractéristiques propres de l'archive, la patrimonialisation du Web ajoute également des questions de représentation au plein sens du terme. Les Suds ou les minorités interrogent une organisation qui, pour être réellement multi-parties prenantes en intégrant notamment une partie de ses utilisateurs dans les processus organisationnels et décisionnels, reste majoritairement dominée par le Nord économique de la planète et ses représentants.

Dans ce contexte spécifique de conservation, l'appropriation de ces nouvelles ressources demande une acculturation à ceux qui souhaitent s'y plonger. Au sein des temporalités

désarticulées de l'archive et de l'océan de données qu'elle représente, de nouvelles méthodes se proposent de mieux armer le regard. La lecture distante ou le travail sur les métadonnées interrogent le fonctionnement des disciplines qui souhaitent intégrer l'archive du Web dans leurs travaux. Ces méthodologies prolongent les questions traditionnelles du rapport entre analyses qualitatives et quantitatives des données de la recherche, mais également celles concernant la montée en compétences du chercheur dont on attend de plus en plus qu'il soit capable de pénétrer les logiques techniciennes de constitution des corpus étudiés.

Enfin, l'archive du Web est l'occasion de repenser en contexte les principaux enjeux éthiques liés à l'oubli, au respect de la vie privée et de la volonté de l'utilisateur, et ce, dans une perspective diachronique. La question de la représentativité dans le domaine de la conservation ou celle du respect de l'anonymat constituent autant d'enjeux déontologiques à débattre dans le cadre d'une pratique collective de l'archive et de construction de la mémoire de nos pratiques en ligne.

C'est sur ces multiples pistes de réflexion que cet ouvrage a souhaité lancer son lecteur, libre à lui de les suivre ou même d'en tracer de nouvelles.

RÉFÉRENCES BIBLIOGRAPHIQUES

ABBATE Janet, 2012, « L'histoire de l'internet au prisme des STS », *Le Temps des médias*, nº 18, p. 170-180.

ANKERSON Megan Sapnar, 2015a, « Read/Write the Digital Archive : Strategies for Historical Web Research », *Digital Research Confidential. The Secrets of Studying Behavior Online*, E. Hargittai, C. Sandvig éd., Cambridge Mass., MIT Press.

ANKERSON Megan Sapnar, 2015b, « Take me back! Web history as chronotourism of the digital archive », *Times and Temporalities of the Web International Symposium*, Paris.

ATKINSON Sarah, WHATLEY Sarah, 2015, « Digital Archives & Open Archival Practices », *Convergence*, vol. 21, nº 1, p. 3-7.

AUBRY Sara, 2010, « Introducing Web Archives as a New Library Service: the Experience of the National Library of France », *LIBER Quarterly. The journal of the Association of European Research Libraries*, [http://persistent-identifier.nl/?identifier=URN:NBN:NL:UI:10-1-113591].

BACHIMONT Bruno, 2017a, « L'archive du Web : une nouvelle herméneutique des traces ? », *Web Corpora*, 21 juin 2017. [https://webcorpora.hypotheses.org/288].

BACHIMONT Bruno, 2017b, *Patrimoine et numérique. Technique et politique de la mémoire*, Bry-sur-Marne, Ina, coll. « Médias et humanités ».

BADOUARD Romain, 2016, « "Je ne suis pas Charlie". Pluralité des prises de parole sur le Web et les réseaux sociaux », *Le défi Charlie. Les médias à l'épreuve des attentats*, P. Lefébure, C. Sécail éd., Paris, Lemieux Éditeurs, [https://hal.archives-ouvertes.fr/hal-01251253/document].

BADOUARD Romain, MABI Clément, MATTOZZI Alvise, SCHUBERT Cornelius, SIRE Guillaume, SORENSEN Estrid, 2016, « STS and media studies: Alternative paths in different countries », *Tecnoscienza*, 7 (1), p. 109-128.

BADOUARD Romain, MUSIANI Francesca, MEADEL Cécile, MONNOYER-SMITH Laurence, 2012, « Towards a Typology of Internet Governance Socio-Technical Arrangements », *Normative Experience in Internet Politics*, F. Massit-Follea, C. Méadel, L. Monnoyer-Smith éd., Paris, Presses des Mines, p. 99-124.

BARATS Christine (dir.), 2013, *Manuel d'analyse du Web en sciences humaines et sociales*, Paris, Armand Colin, coll. « U Sciences humaines et sociales ».

BAZIN Maëlle, 2017, « Quand la rue prend le deuil. Les mémoriaux éphémères après les attentats », laviedesidees.fr, 26 mai, [http://www.laviedesidees.fr/IMG/pdf/20170526_bazinmemo-2.pdf].

BEAUDOIN Valérie, PEHLIVAN Zeynep, 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". [Rapport de recherche] », Bibliothèque nationale de France ; Bibliothèque de documentation internationale contemporaine ; Télécom ParisTech, [https://hal.archives-ouvertes.fr/hal-01425600].

BEN-DAVID Anat, 2016 « What does the Web Remember of its Deleted Past? An Archival Reconstruction of the Former Yugoslav Top Level Domain », *New Media & Society*, 18(7), p. 1103–1119.

BEN-DAVID Anat, AMRAM Adam, BEKKERMAN Ron, 2016, « The colors of the national Web: visual data analysis of the historical Yugoslav Web domain », *International Journal on Digital Libraries*, [http://www.academia.edu/30508702/The_colors_of_the_national_web_Visual_data_analysis_of_the_historical_Yugoslav_web_domain].

BEN-DAVID Anat, AMRAM Adam, 2018, « The Internet Archive and the socio-technical construction of historical facts », *Internet Histories*, DOI : 10.1080/24701475.2018.1455412.

BEN-DAVID Anat, HUURDEMAN Hugo, 2014, « Web Archive Search as Research: Methodological and Theoretical Implications », *Alexandria: The Journal of National and International Library and Information Issues*, nº 25, p. 93-111.

BORTZMEYER Gabriel, 2016, « Révolution, touche replay », *Vacarme*, 77, p. 74-83.

BOULLIER Dominique, 2015, « Charlie est un phénomène de 3e génération (aussi) », SHS 3G, [http://shs3g.hypotheses.org/114].

BOUVIER Yves, POLINO Marie-Noëlle, VARASCHIN Denis, 2010, « Introduction. Patrimoine de la communication des entreprises de réseau », *Flux*, 82 (4), p. 5-7.

BRÜGGER Niels, 2009, « Website history and the website as an object of study », *New Media & Society*, vol. 11, nº 1-2, p. 115-132.

BRÜGGER Niels, 2011 « Web archiving—Between past, present, and future », *The handbook of Internet studies*, p. 24-42.

BRÜGGER Niels, 2012a, « L'historiographie de sites web : quelques enjeux fondamentaux », *Le Temps des médias*, nº 18/1, p. 159-169.

BRÜGGER Niels, 2012b, « Web History and the Web as a Historical Source », *Zeithistorische Forschungen/Studies in Contemporary History*, 9, p. 316-325, [http://www.zeithistorische-forschungen.de/2-2012/id%3D4426].

BRÜGGER Niels, 2015, *Humanities, Digital Humanities, Media studies, Internet studies: An inaugural lecture*, Aarhus, The Centre for Internet Studies, Monograph Series, 16.

BRÜGGER Niels, 2016, « Digital Humanities in the 21st Century: Digital Material as a Driving Force », *Digital Humanities Quaterly*, preview, vol. 10, nº 2, [http://www.digitalhumanities.org/dhq/vol/10/3/000256/000256.html].

BRÜGGER Niels, 2017, « Probing a nation's web domain: A new approach to web history and a new kind of historical source », *The Routledge Companion to Global Internet Histories*, G. Goggin, M. McLelland éd., New York/Abingdon, Routledge, p. 61-73.

BRÜGGER Niels, LAURSEN Ditte, NIELSEN Janne, 2017, « Exploring the domain names of the Danish web », *The Web as History*, N. Brügger, R. Schroeder éd., Londres, UCL Press, p. 238-248.

BRUNTON Finn, 2017, « Notes from/dev/null », *Internet Histories*, nº 1-2, p. 138-145.

BUTLER Chris, 2017, « Who Blocked the Archive in Jordan? », *Internet Archive Blogs*, [https://blog.archive.org/2017/04/11/who-blocked-the-archive-in-jordan/].

BYGRAVE Lee, BING Jon, 2009, *Internet Governance. Infrastructure and Institutions*, Oxford, Oxford University Press.

CHABIN Marie-Anne, 2011, « Peut-on parler de diplomatique numérique », *Le Blog de Marie-Anne Chabin*, [http://www.marieannechabin.fr/diplomatique-numerique/].

CHABIN, Marie-Anne, 2012 « L'ère numérique du faux », *Médium*, n° 31/2, p. 46-66.

CHAIMBAULT Thomas, 2008, « L'archivage du Web », Dossier documentaire [en ligne], Villeurbanne, Enssib, [http://www.enssib.fr/bibliotheque-numerique/documents/1730-l-archivage-du-web.pdf].

CHENITI Tarek, 2009, *Global Internet Governance in Practice. Mundane Encounters and Multiple Enactments*, thèse de doctorat, University of Oxford.

CHEVALLIER Philippe, ILLIEN Gildas, 2011, « Les archives de l'internet. Une étude prospective sur les représentations et les attentes des utilisateurs potentiels », Rapport BnF, [http://www.bnf.fr/documents/enquete_archives_web.pdf].

CHUN Wendy Hui Kyong, 2011, *Programmed visions: Software and memory*, Cambridge Mass., MIT Press.

CLAVERT Frédéric, 2018a, « Le goût de l'API », *Le goût de l'archive à l'ère numérique*, F. Clavert, C. Muller éd., [http://www.gout-numerique.net/table-of-contents/gout-api].

CLAVERT Frédéric, 2018b, « Temporalités du centenaire de la Grande Guerre sur Twitter », *Temps et temporalités du Web*, V. Schafer éd., Nanterre, Presses de Nanterre, p. 113-134.

COHEN Évelyne, VERLAINE Julie, 2013, « Le dépôt légal de l'internet français à la Bibliothèque nationale de France », *Sociétés & Représentations* , vol.1, n° 35, p. 209-218.

COSTEA Maria-Dorina, 2018, « Report on the Scholarly Use of Web Archive », Aarhus, Netlab, [http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf].

DENARDIS Laura, 2014, *The Global War for Internet Governance*, New Haven, Yale University Press.

DIMINESCU Dana, LOVELUCK Benjamin, 2014, « Traces of dispersion: online media and diasporic identities. Crossings: Journal of Migration & Culture », *Intellect*, 5 (1), p. 23-39.

DOUGHERTY Meghan, 2017, « Wayfinder: Building an interface for a Web archive », IWAW'07 [en ligne], [https://www.researchgate.net/publication/228938220_Wayfinder_Building_an_Interface_for_a_Web_Archive].

DOUGHERTY Meghan, MEYER Eric T., MADSEN McCARTHY Christine *et al.*, 2010, *Researcher Engagement with Web Archives: State of the Art*, Londres, JISC.

DULONG DE ROSNAY Mélanie, GUADAMUZ Andrés, 2017, « Memory Hole or Right to Delist? Implications of the Right to Be Forgotten for Web Archiving », *RESET*, 6, 2017, [http://journals.openedition.org/reset/807].

FINNEMAN Niels Ole, 2015, « Hypertextual relations in digital born materials. Hypertext and time: Towards a genre analysis of heterogeneous digital materials », *RESAW Conference*, Université d'Aarhus, Danemark, 8 au 10 juin.

GALISON Peter, 1997, *Image & logic : A material culture of microphysics*, Chicago, The University of Chicago Press.

GAME Valérie, ILLIEN Gildas, 2006, « Le dépôt légal d'Internet à la Bibliothèque nationale de France : cadre juridique, modèle de collecte, évolution des métiers », *Bulletin des bibliothèques de France* [en ligne], n° 3, p. 82-85.

GEBEIL Sophie, 2014, « Pourquoi archiver le Web ? Les missions de l'IIPC », *Carnet de recherche Internet, histoire et mémoires*, [https://madi.hypotheses.org/243].

GEBEIL Sophie, 2015, « La fabrique numérique des mémoires de l'immigration maghrébine sur le Web français (1999-2014) », thèse de doctorat en histoire sous la direction de Maryline Crivello, Aix-en-Provence, Université Aix-Marseille.

GEBEIL Sophie, 2016, « Quand l'historien rencontre les archives du Web », *Revue de la BNF*, 53 (2), p. 185-191.

GEBEIL Sophie, 2017, « La patrimonialisation numérique des mémoires de l'immigration maghrébine en France dans les années 2000 », *RESET* [en ligne], 6, 2017, mis en ligne le 30 octobre 2016, [http://journals.openedition.org/reset/853].

GIGLIETTO Fabio, LEE Yenn, 2015, « To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 Charlie Hebdo Shooting in France », #Microposts2015 · 5th Workshop on Making Sense of Microposts @WWW2015, [http://ceur-ws.org/Vol-1395/paper_12.pdf].

GOMES Daniel, MIRANDA João, COSTA Miguel, 2011, « A Survey on Web Archiving Initiatives », *TPDL 2011. Proceedings of the 15th international conference on Theory and practice of digital libraries: research and advanced technology for digital libraries*, Berlin/Heidelberg/New York, Springer, p. 408-420.

GRAHAM Shawn, MILLIGAN Ian, WEINGART Scott, 2015, *Exploring Big Historical Data. The Historian's Macroscope*, Londres, Imperial College Press.

HALE Scott, BLANK Grant, ALEXANDER Victoria, 2017, « Live versus archive: Comparing a web archive to a population of web pages », *The Web as History*, N. Brügger, R. Schroeder éd., Londres, UCL Press, p. 45-61.

HUUDERMAN Hugo *et al.*, 2015, « Lost but not forgotten: finding pages on the unarchived web », *International Journal on Digital Libraries*, 16, 3, p. 247-265.

ILLIEN Gildas, 2011, « Une histoire politique de l'archivage du web ». *Bulletin des bibliothèques de France (BBF)*, n° 2, p. 60-68, [http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012].

JONES Meg Leta, 2016, *Ctrl+ Z : The right to be forgotten?* New York, NYU Press.

KAHLE Brewster, 1997, « Preserving the Internet », *American Scientific*, n° 276, p. 82-83.

KAHLE Brewster, 2011, « Internet Archive, Le meilleur du web est déjà perdu », [http://www.internetactu.net/2011/06/28/brewster-kahle-internet-archive-le-meilleur-du-web-est-deja-perdu/].

KAHLE Brewster, 2014a, « Help Us Keep the Archive Free, Accessible, and Reader Private », *Internet Archive Blog*, [https://blog.archive.org/2016/11/29/help-us-keep-the-archive-free-accessible-and-private/].

KAHLE Brewster, 2014b, « Please Help Protect Net Neutrality », *Internet Archive Blog*, [https://blog.archive.org/2014/09/10/please-help-protect-net-neutrality/].

KESSOUS Emmanuel, 2012, *L'attention au monde : Sociologie des données personnelles à l'ère numérique*, Paris, Armand Colin.

KIRSCHENBAUM Matthew, OVENDEN Richard, REDWINE Gabriela, 2010, *Digital forensics and born-digital content in cultural heritage collections*, Washington, D.C, Council on Library and Information Resources.

LATZKO-TOTH Guillaume, PASTINELLI Madeleine, 2013, « Par-delà la dichotomie public/privé : la mise en visibilité des pratiques numériques et ses enjeux éthiques », *tic&société* [nn ligne], vol. 7, nº 2, [http://journals.openedition.org/ticetsociete/1591].

LATZKO-TOTH Guillaume, PROULX Serge, 2013, « Enjeux éthiques de la recherche en ligne », *Manuel d'analyse du Web en sciences humaines et sociales*, C. Barats éd., Paris, Armand Colin.

LOBBÉ Quentin, 2018, *Archives, fragments web et diasporas. Pour une exploration désagrégée de corpus d'archives web liées aux représentations en ligne des diasporas*, thèse de doctorat de l'université Paris-Saclay, sous la direction de Pierre Senellart et Dana Diminescu, Paris.

MABI Clément, PLANTIN Jean-Christophe, MONNOYER-SMITH Laurence, 2014, « Interroger les données en SHS à partir de leur écosystème », *Information et communication scientifiques à l'heure du numérique*, V. Schafer éd., Paris, CNRS Éditions, coll. « Les essentiels d'Hermès », p. 63-78.

MARRES Noortje, 2012, « The redistribution of methods: on intervention in digital social research, broadly conceived », *The Sociological Review*, 60, p. 139-165.

MARRES Noortje, GERLITZ Carolin, 2015, « Interface methods: renegotiating relations between digital social research, STS and sociology », *The Sociological Review*, 64 (1), p. 21-46.

MASANÈS Julien (dir.), 2006, *Web Archiving*, Berlin/Heidelberg, Springer.

MERZEAU Louise, 2012, « Faire mémoire de nos traces numériques », E-dossiers de l'audiovisuel, [https://halshs.archives-ouvertes.fr/halshs-00727308. halshs-00727308].

MERZEAU Louise, 2014, « Vers un Web temporel », conférence à la IIPC General Assembly, [http://merzeau.net/vers-un-web-temporel/].

MERZEAU Louise, MUSSOU Claude, 2017, « L'expérience des ateliers du dépôt légal du Web de l'Ina », Carnet de recherche WebCorpora, [http://webcorpora.hypotheses.org/302].

MILLIGAN Ian, 2012, « Mining the "Internet Graveyard": rethinking the historians' toolkit », *J. Can. Hist. Assoc. Revue de la Société historique du Canada*, 23 (2), p. 21-64.

MILLIGAN Ian, 2016, « Lost in the Infinite Archive: The Promise and Pitfalls of Web Archives », *International Journal of Humanities and Arts Computing*, vol. 10 n°1, p. 78-94, ISSN 1753-8548, [https://www.euppublishing.com/doi/full/10.3366/ijhac.2016.0161].

MILLIGAN Ian, 2017, « Welcome to the Web: The online community of GeoCities during the early years of the World Wide Web », *The Web as History*, N. Brügger, R. Schroeder éd., Londres, UCL Press, p. 137-157.

MILLIGAN Ian, 2018, « Ethics and the Archived Web Presentation: "The Ethics of Studying GeoCities"» , Ian Milligan's blog, [https://ianmilligan.ca/2018/03/27/ethics-and-the-archived-web-presentation-the-ethics-of-studying-geocities/].

MOIRAGHI Eleonora, 2018, « Le projet Corpus et ses publics potentiels. Une étude prospective sur les besoins et les attentes des futurs usagers », Paris, BnF, [https://hal-bnf.archives-ouvertes.fr/hal-01739730/document].

MORETTI Franco, 2013, *Distant Reading*, Londres/New York, Verso.

MOROZOV Evgeny, 2014, *Pour tout résoudre cliquez ici. L'aberration du solutionnisme technologique*, Paris, FYP Éditions.

MUSIANI Francesca, SCHAFER Valérie (éd.), 2017, « Patrimoine et patrimonialisation numériques », dossier *RESET*, n° 6.

MUSIANI Francesca, SCHAFER Valérie, 2019, « Science and Technology Studies Approaches to Web History », *The Handbook of Web History*, N. Brügger, I. Milligan éd., Londres, Sage.

MUSSO Marta, MERLETTI Francesco, 2016, « This is the future: A reconstruction of the UK business web space (1996-2001) », *New Media & Society*, vol 18, n° 7, p. 1120-1142.

MUSSOU Claude, 2012, « Et le Web devint archive : enjeux et défis », *Le Temps des médias*, vol. 19, nº 2, p. 259-266.

NANNI Federico, 2017, « Reconstructing a website's lost past – Methodological issues concerning the history of www.unibo.it », *Digital Humanities Quarterly*, [http://www.digitalhumanities.org/dhq/vol/11/2/000292/000292.html].

NATALE Enrico, 2017, « Les médiations numériques du patrimoine », *RESET* [en ligne], 6, 2017, mis en ligne le 30 octobre 2016, [http://journals.openedition.org/reset/787].

NIU Jinfang, 2012, « An overview of web archiving », *D-Lib magazine*, vol. 18, nº 3-4.

NOIRET Serge, 2011, « Y a-t-il une histoire numérique 2.0 », *Les historiens et l'informatique : Un métier à réinventer. Études réunies par Jean-Philippe Genet et Andrea Zorzi*, J-P. Genet, A. Zorzi éd., Rome, École française de Rome, coll. de l'École française de Rome, p. 235-288, [http://cadmus.eui.eu/handle/1814/18074].

NORA Pierre, 1996, « Préface », *La France du patrimoine. Les choix de la mémoire*, M.-A. Sire éd., Paris, Gallimard/MONUM.

OURY Clément, 2012, « Soixante millions de fichiers pour un scrutin. Les collections de sites politiques à la BNF », *Revue de la BNF*, vol. 40, nº. 1, p. 84-90, [https://www.cairn.info/revue-de-la-bibliotheque-nationale-de-france-2012-1-page-84.htm].

PABON CADAVID JHONNY Antonio, SATHIK BASHA Johnkhan, KALEESWARAN Gandhimani, 2013, « Legal and Technical Difficulties of Web Archival in Singapore », *NTU*, [http://library.ifla.org/217/1/198-cadavid-en.pdf].

PALOQUE-BERGÈS Camille, 2014, « Le rôle des communautés patrimoniales d'Internet dans la constitution d'un patrimoine numérique : des mobilisations diverses autour de l'auto-médiation », *Heritage and Digital humanities*, B. Dufrêne, B. Barbier éd., Berlin, Lit.Verlag, p. 277-290.

PALOQUE-BERGÈS Camille, 2017, « Usenet as a web archive. Multi-layered archives of computer-mediated communication », *Web 25: histories from the first 25 years of the World Wide Web*, N. Brügger éd., Londes, Peter Lang, p. 227-250.

PALOQUE-BERGÈS Camille, 2018, *Qu'est-ce qu'un forum internet ? Une généalogie historique au prisme des cultures savantes numériques*, Nouvelle édition [en ligne], Marseille, OpenEdition Press, ISBN : 9791036500350. DOI : 10.4000/books.oep.1843, [http://books.openedition.org/oep/1843].

PALOQUE-BERGÈS Camille, SCHAFER Valérie, 2015, « Quand la communication devient patrimoine », *Hermès*, n° 71, p. 255-262.

PAPY Fabrice, 2015, *Bibliothèques numériques. Interopérabilité et usages*, Paris, ISTE Éditions.

PARIKKA Jussi, 2013, « Archival Media Theory: An Introduction to Wolfgang's Ernst Media Archeology », *Digital Memory and the Archive*, W. Ernst éd., Minneapolis, University of Minnesota Press, p. 1-22.

PLANTIN Jean-Christophe, MONNOYER-SMITH Laurence, 2013, « Ouvrir la boîte à outils de la recherche numérique. Trois cas de redistribution de méthodes », *tic&société*, vol. 7, n° 2, [https://ticetsociete.revues.org/1527].

PROULX Serge, 2009, « L'intelligence du grand nombre : la puissance d'agir des contributeurs sur Internet – limites et possibilités », *7e colloque du chapitre français de l'ISKO, Intelligence collective et organisation des connaissances*, Lyon, 24-26 juin 2009, [http://pro.ovh.net/~iskofran/pdf/isko2009/PROULX.pdf].

RIEDER Bernard, RHÖLE Theo, 2012, « Digital Methods: Five Challenges », *Understanding Digital Humanities*, D. Berry éd., New York, Palgrave Macmillan, p. 67-84.

ROUSTAN Mélanie (dir.), 2016, *La recherche dans les institutions patrimoniales : sources matérielles et ressources numériques*, Villeurbanne, Presses de l'Enssib.

SCHAFER Valérie, 2015, *En construction. Une histoire française du Web des années 1990*, HDR, vol. 2, Université Paris-Sorbonne, Paris.

SCHAFER Valérie, 2017, « Archives : comment le Web devient patrimoine », *The Conversation*, 24 avril, [https://theconversation.com/archives-comment-le-web-devient-patrimoine-76487].

SCHAFER Valérie, 2019, « Exploring the "French Web" of the 90s'», *The Historical Web and Digital Humanities: the Case of National Web domains*, N. Brügger, D. Laursen éd., New York/Abingdon, Routledge.

SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, 2016, « Negotiating the Web of the Past. Web archiving, governance and STS », *French Journal for Media Research* [en ligne], 6, numéro spécial *La toile négociée / Negotiating the Web*, [http://frenchjournalformediaresearch.com/lodel/index.php?id=952].

SCHAFER Valérie, MUSIANI Francesca, BORELLI Marguerite, 2017, « Le patrimoine culturel immatériel pour aider à penser le patrimoine activement numérique », *Patrimoine culturel immatériel et numérique*, M. Sévero, S. Cachat éd., Paris, L'Harmattan, p. 131-145.

SEVERO Marta, CACHAT Séverine (dir.), 2017, *Patrimoine culturel immatériel et numérique*, Paris, L'Harmattan.

SIRE Guillaume, 2015, « Inclusion exclue : le code est un contrat léonin », *Réseaux*, vol. 189/1, p. 187-214.

SMIT Eefke, VAN DER HOEVEN Jeffrey, GIARETTA, David, 2011, « Avoiding a Digital Dark Age for data: why publishers should care about digital preservation », *Learned Publishing*, vol. 24, nº 1, p. 35-49.

STAR Susan Leigh, GRIESEMER James, 1989, « Institutional ecology, "translations" and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39 » *Social Studies of Science*, 19 (3), p. 387-420.

TAYLOR Nicholas, 2015, « Questions of ethics at Web Archives 2015 », Stanford Libraries website, [http://library.stanford.edu/blogs/digital-library-blog/2015/12/questions-ethics-web-archives-2015].

THOËR Christine, MILLERAND Florence, MYLES David, ORANGE Valérie, GIGNAC Olivier, 2012, « Enjeux éthiques de la recherche sur les forums internet portant sur l'utilisation des médicaments à des fins non médicales », *Communiquer*, 7, p. 1-22.

TRELEANI Matteo, 2017, *Qu'est-ce que le patrimoine numérique ? Une sémiologie de la circulation des archives*, Lormont, Le Bord de l'eau.

TROMPETTE Pascale, VINCK Dominique, 2009, « Retour sur la notion d'objet-frontière », *Revue d'anthropologie des connaissances*, vol. 3, nº 1, p. 5-27.

UNESCO, 2012, « The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage », Conference Memory of the World 20th Anniversary, September 26-28th, Vancouver, Canada.

VERRY Élisabeth (éd.), 2007, « Archives et Internet : contributions et témoignages », dossier de *La Gazette des archives*, n° 207/3.

WEBSTER Peter, 2018, « Religion in Web history: a survey », *The Sage Handbook to Web History*, N. Brügger, I. Milligan éd., Londres, Sage.

WINNER Langdon, 1980, « Do artifacts have politics? », *Daedalus*, vol. 109, n°1, p. 121-136.

WINTERS Jane, 2017, « Coda: Web archives for humanities research – some reflections », *The Web as History*, N. Brügger, R. Schroeder éd., Londres, UCL Press, p. 238-248.

WINTERS Jane, 2017a, « A breakthrough year for web archiving in 2016? », dpconline.org, 3 février 2017, [http://dpconline.org/blog/a-breakthrough-year-for-web-archiving-in-2016].

WINTERS Jane, 2017b, « Breaking into the mainstream: demonstrating the value of internet (and web) histories », *Internet histories*, vol. 1, n° 1-2, p. 173-179.

WORKING GROUP ON INTERNET GOVERNANCE, 2005, *Report of the WGIG*, Château de Bossey, juin 2005, [https://www.wgig.org/docs/WGIGREPORT.pdf].

Webographie

Internet Archive [https://archive.org/index.php]

Archive Team [https://archiveteam.org]

Archive-It [https://archive-it.org]

RESAW [http://resaw.eu]

Site du projet ASAP (Archives sauvegarde attentats Paris) soutenu par le CNRS en 2016 [http://asap.hypotheses.org]

Site du projet Web90 soutenu par l'ANR de 2014 à 2018 [http://web90.hypotheses.org]

Blog Web Archives for Historians (Ian Milligan et Peter Webster) [https://webarchivehistorians.org]

Site de l'IIPC (International Internet Preservation Consortium) [http://netpreserve.org]

Carnet Web Corpora. Explorer les archives de l'internet à la BnF [http://webcorpora.hypotheses.org]

Site des ateliers du DL Web Ina [http://atelier-dlweb.fr/blog/]

LES AUTEURS

Francesca Musiani est chargée de recherche au CNRS. Elle est également chercheuse associée au Centre de sociologie de l'innovation (i3/Mines ParisTech) et Global Scholar auprès de l'Internet Governance Lab de l'American University (Washington DC, USA). Ses travaux portent sur la gouvernance de l'Internet. Elle a actuellement des rôles de coordination dans des projets de recherche qui portent sur le chiffrement et les « résistances numériques ».

Camille Paloque-Bergès est docteure en Sciences de l'information et de la communication. Elle étudie l'histoire technique, sociale et culturelle des réseaux informatiques. Elle est actuellement ingénieure de recherche au laboratoire HT2S du CNAM et a publié, dans la même collection « Encyclopédie Numérique », en 2018 : *Qu'est-ce qu'un forum Internet ? Une généalogie historique au prisme des cultures savantes numériques*.

Valérie Schafer est professeur d'histoire européenne contemporaine au C2DH à l'université du Luxembourg. Ses recherches portent sur l'histoire des télécommunications et de l'informatique, notamment des réseaux de données et du numérique. Elle a coordonné au CNRS de 2014 à début 2018 le projet Web90, soutenu par l'Agence Nationale de la Recherche, auquel ont participé tous les auteurs de ce livre.

Benjamin G. Thierry est maître de conférences en histoire contemporaine à Sorbonne Université. Spécialiste en histoire de l'informatique et des télécommunications, il porte un intérêt particulier aux processus et aux vecteurs de diffusion du numérique au sein de la société (interfaces homme-machine, processus organisationnels et éducation au numérique). Il est également vice-président numérique de Sorbonne Université.

TABLE DES MATIÈRES