Archivage du Web, un enjeu de gouvernance (d'Internet)

Francesca Musiani

Archivage du Web, un enjeu de gouvernance (d'Internet)

Cet article est une version auteur de :

Francesca Musiani, 2022, « Archivage du Web, un enjeu de gouvernance (d'Internet) », in Clarisse Bardiot, Esther Dehoux, Emilien Ruiz (dir.) *La fabrique numérique des corpus en sciences humaines et sociales*, Presses Universitaires du Septentrion.

Auteur

Francesca Musiani, docteure en socio-économie de l'innovation de MINES ParisTech (2012), est chargée de recherche au CNRS et directrice adjointe du Centre Internet et Société du CNRS (UPR 2000). Elle est chercheuse associée au Centre de sociologie de l'innovation (i3/MINES ParisTech : http://www.csi.mines-paristech.fr/) et *Global Fellow* auprès de l'Internet Governance Lab (http://internetgovernancelab.org/). de l'American University à Washington, DC. Les travaux de Francesca Musiani portent sur la gouvernance de l'Internet, dans une perspective interdisciplinaire.

Résumé

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 ?* »). 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. Les points suivants seront abordés :

1. 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 ; et enfin le secteur privé.
2. L'archivage du Web n'échappe 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.
3. L'archivage du Web révèle également la présence de tensions géopolitiques, et on y retrouve des dynamiques qui rappellent le problème de la fracture numérique.
4. Enfin, 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.

Introduction

La plupart des institutions de collecte des archives du Web livrent en ligne un aperçu de leurs 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 aussi 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[2].

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 aura 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 socio-numériques renouvellent aussi les questions de sélection et de capture. Autant d'éléments liés à la problématique de la gouvernance des archives du Web qui se pose dès lors comme un microcosme de questions plus larges de gouvernance d'Internet[3].

Des archivages en constante évolution

Une archive du Web est loin d'être un objet statique[4] : elle évolue sous l'effet des modalités de collecte, de la profondeur de l'exploration, ainsi que des changements au fil du temps de caractéristiques 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 plus mémorielle, accompagnée de méthodes de « bricolage » reflétant le manque d'une meilleure alternative. Une deuxième 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. 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

1 Voir sur le site de la BNF : http://www.bnf.fr/fr/professionnels/archivage_web_bnf/a.dlweb_collecte_acces_libre.html.

2 Voir (Ben-David et Amram 2018) sur le Web archivé nord-coréen.

3 Ce chapitre se fonde sur « Où commence et s'arrête l'archive ? » du livre *Qu'est-ce qu'une archive du Web ?* que l'auteur a publié début 2019 avec Valérie Schafer, Camille Paloque-Bergès et Benjamin Thierry (Marseille, OpenEdition), disponible en accès libre intégral : https://books.openedition.org/oep/8713.

4 Cette section reprend des éléments de (Schafer, Musiani et Borelli 2016).

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 considérée comme une caractéristique essentielle :

> 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)

Avec cette attention particulière prêtée aux variations du Web « vivant », le Web archivé s’éloigne progressivement de l’idée d’une restitution. 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. Celui-ci devient partie intégrante de l’intelligibilité des contenus car il est inscrit dans sa propre époque et 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 qu’on n’est presque jamais en train de, tout simplement, « faire une copie ». Le processus d’archivage du Web crée une série de versions uniques d’un contenu, où quelque chose peut être perdu et autre chose, qui n’était pas en ligne à cet instant T, peut être archivé avec ce contenu. Ce qui peut rendre très difficile 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 2015).

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 – ce qui explique pourquoi les utilisateurs se trouvent régulièrement face à des pages web manquantes ou non trouvées. Cela répond à l’effort pour 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 eu 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).

Les polices et caractères peuvent aussi différer dans les archives du Web par rapport aux pages originelles ; si au moment de l'archivage une 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 peut poser problème dans ce paysage mouvant : plusieurs pages web des années 1990, désormais archivées, montrent des trous béants là où leurs images étaient autrefois. Probablement, la raison à la base de ce phénomène est à rechercher moins dans la difficulté technique de la capture, et plus 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, et afin d'éviter les doublons, celles-ci ne sont pas systématiquement re-collectées ; ainsi, si leur URL n'a pas changé d'un *crawl* à l'autre, elles peuvent être récupérées du *crawl* le plus récent, au lieu d'être à nouveau capturées. Cela explique également 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.

Le périmètre de l'archive du Web

C'est le regard que l'on porte sur l'archive qui, 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 site web a donné lieu à des réflexions méthodologiques et épistémologiques[5], mais qui tendent 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…). Ainsi, la multitude des niveaux, formats et éléments documentaires concernés par l'archivage (textes, images, sons, vidéo, graphismes, bases de données, logiciels, codes…) entre dans un périmètre plus ou moins cohérent selon la manière dont on l'analyse.

Toutefois, le regard du chercheur est *in fine* 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, définit 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, passant par :

1. L'évaluation et la sélection, qui même dans le cas de collections non discriminantes des contenus se font sur la base de postulats et de critères au moins sous-jacents. Par exemple, pour Internet Archive qui *a priori* ne trie pas sa récolte, c'est

5 Voir par exemple (Barats 2013).

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, recommandation…). 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 Taiwan a recours à l'échantillonnage.

2. L'acquisition : si la tradition institutionnelle de dons et 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).
3. 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).
4. 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…), ou bien induites à partir d'une extraction des métadonnées du code des pages d'origine.
5. 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 buts juridiques ou scientifiques ; 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…).

La question de la création des collections révèle nombre d'autres enjeux sous-jacents. Par exemple, comment rendre la cohérence d'une collection à partir de nouveaux genres éditoriaux nativement numériques, encore mal identifiés ? Les collections de blogs ont ainsi retenu l'attention, pour les problèmes qu'ils posent en matière 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 amateure de contenus de médias en ligne, 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].

On remarque que de nombreuses contraintes limitent le périmètre des archives du Web telles qu'elles s'offrent à l'usager. 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. 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 crée 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 »[7] 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. Sarah Atkinson et Sarah Whatley (2015) rappellent ainsi que les archives numériques doivent être mises en perspective avec l'espace public numérique. Ainsi, 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.

L'archivage des réseaux socio-numériques, 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 socio-numériques (RSN). Dispositifs de flux, dont Frédéric Clavert (2017) 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 URL parfois réduites, etc. Les contenus des RSN ne sont donc pas toujours aisément accessibles ou archivables, 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 de 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[8] pour qu'il capture bien

6 Pour en savoir plus, consulter : https://cordis.europa.eu/project/rcn/98063_fr.html.

7 Kuchler, Hannah. 2014. « How to preserve the web's past for the future ». *Financial Times*, 11 avril 2014. https://www.ft.com/content/d87a33d8-c0a0-11e3-8578-00144feabdc0.

les contenus qui nous intéressent. En particulier, les protocoles HTTPS[9] nous posent parfois des problèmes, tout comme Facebook lorsqu'il utilisait des *captcha* »[10]. 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 ensuite 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[11].

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[12]», 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 pose également la question des modalités concrètes d'accès à ces archives : leur accessibilité pour des chercheurs par exemple européens impliquera-t-elle de 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 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[13].

De plus, comme pour le Web, le périmètre de collecte est aussi sélectif pour les RSN et si l'INA a pris la mesure de l'intérêt de l'archivage de Twitter et lancé des collectes dès 2014,

---

8 Robot d'indexation utilisé par la BNF mais aussi par Internet Archive. https://webarchive.jira.com/wiki/spaces/Heritrix.

9 Protocole web sécurisé.

10 Entretien mené par Marguerite Borelli et Valérie Schafer dans le cadre du projet *ASAP*, le 21 mars 2016. https://asap.hypotheses.org/168.

11 Voir l'article de Plaugic, Lizzie. 2017. « The Library of Congress will no longer archive every tweet ». *The Verge*, 26 décembre 2017. https://www.theverge.com/2017/12/26/16819748/library-of-congress-twitter-archive-project-stalled.

12 D'après « Update on the Twitter Archive at the Library of Congress ». 2017. Library of Congress. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

13 Pour un aperçu des chiffres, voir Coëffé, Thomas. 2017. « Les 50 chiffres à connaître sur les médias sociaux en 2018 ». *BDM* (blog). 28 décembre 2017. https://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2018/.

l'équipe dédiée au DL Web (pour Dépôt Légal 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*.

Mais son expérience s'est aussi manifestée lors des attentats de 2015, au moment où des millions de tweets ont réagi aux évènements autour de *Charlie Hebdo* puis à ceux de novembre 2015, suscitant aussi la réactivité de chercheurs qui lancent des collectes très rapidement (par exemple la collecte de Romain Badouard qui sert de base à sa réflexion sur le « Je ne suis pas Charlie » (Badouard 2016), celle du canadien Nick Ruest, dont les données sont accessibles en ligne[14], ou encore celles de Giglietto et Lee (2015).

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 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] 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 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[15].

L'INA a pleinement conscience de l'intérêt de démarrer la collecte tôt, 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 », ajoute Zeynep Pehlivan[16].

Si l'aspect des archives du Web peut changer d'une institution à une autre, le cas de Twitter est particulièrement révélateur : la BNF utilise le robot de capture Heritrix développé par Internet Archive 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. Les deux interfaces de programmation, API Search et Streaming, permettant pour la première à un utilisateur de remonter à un contenu particulier sur les sept derniers jours, et pour la seconde de capter un flux au fur et à mesure pour une requête précise, sont gratuites et publiques. Il est aussi possible de récupérer *a posteriori* les données de Twitter de façon payante. L'API publique a des limites : on ne 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 évènements 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 joue en

14 Voir : https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10864/10830.

15 Entretien réalisé par Valérie Schafer fin 2017.

16 *Idem*.

urgence. D'autres biais ou limites ne peuvent être ignorés du chercheur : par exemple le nombre de retweets (re-publication de tweet 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 : 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* »[17] (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 la même année l'affaire des articles disparus de la cour suprême américaine révélée par *The New York Times*[18] – on parle alors de « *reference rot* ». Les liens et contenus web s'effacent 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 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 une telle encapsulation de contenu a été l'un des premiers obstacles au moissonnage de données web par l'outil 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 re-publication ou de l'administration de forums Internet dont la vie peut être déterminée par des protocoles différents de ceux du Web : mal gérés par leurs administrateurs, difficiles à naviguer, impossibles à collecter, ils tendent à devenir des « ruines numériques » sur le Web (Paloque-Bergès 2017, 2018).

---

17 Summers, Ed. 2015. « The Web as a Preservation Medium ». *On Archivy* (blog). 7 mai 2015. https://medium.com/on-archivy/the-web-as-a-preservation-medium-3d697328b3b8.

18 Liptak, Adam. 2013. « In Supreme Court Opinions, Web Links to Nowhere ». *The New York Times*, 23 septembre 2013. https://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html.

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 protocole 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).

Toutefois, comme l'analyse Guillaume Sire, ce contrat de code repose sur un consensus léonin, c'est-à-dire régit 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).

« *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 don't change* ». Des méthodes alternatives émergent pour pallier les difficultés des archivistes numériques. 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). Ces méthodes viennent tenter de répondre aux interrogations traditionnelles que les historiens renouvellent face aux archives numériques : 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.

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 ?*). Avec ce mot « politique », Winner mettait à l'épreuve la question de la neutralité technologique, pour rechercher dans 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). Si on cherche à appliquer cette hypothèse à l'étude des archives du Web, il s'agit de comprendre les manières dont la nature distribuée, et inscrite dans la technique, de l'archivage du Web lui permet d'incarner des formes spécifiques d'autorité et de pouvoir (DeNardis 2014), ce qui ferait de ce domaine un microcosme de la gouvernance de l'Internet au sens plus large. Cette démarche a occupé certains de nos travaux récents (Schafer, Musiani et Borelli 2016 ; Musiani et Schafer 2019)[19].

L'archivage du Web repose sur un modèle multi-parties prenantes. Une variété d'acteurs sont concernés par l'archivage du Web : des fondations comme Internet Archive ; des

19 Sur lesquels cette section se base.

organisations transnationales, à commencer par l'IIPC ; la société civile (des militants de l'Archive Team à d'autres initiatives fondées par des communautés de chercheurs) ; et enfin, le secteur privé (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[20]). Ainsi, on retrouve dans l'archivage du Web les principales catégories d'acteurs impliqués dans la gouvernance d'Internet, ainsi que leurs tensions et leurs alliances. Des expériences de collaboration entre des institutions d'archivage et des équipes de recherche voient ainsi régulièrement le jour ; la BNF a par exemple associé notre équipe *Web90*[21] à une réflexion sur l'implémentation du 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. L'Internet Archive va encore plus loin en promouvant explicitement des initiatives *bottom-up* destinées à revaloriser l'intervention humaine dans un monde où « les machines allaient nous sauver – parcourant le Web, numérisant les ouvrages, organisant l'information [mais ce sont] les communautés de gens qui sont au centre de l'archivage[23] » (Kahle 2014).

L'archivage du Web n'échappe pas à des tensions ayant trait à la standardisation, un des enjeux traditionnellement le plus vif de la gouvernance d'Internet, et à des imaginaires et visions divergents, des communs aux formats propriétaires. À ce propos, il est souvent intéressant d'observer les types et les périmètres des organisations d'archivage du Web. Depuis août 2006, la mission de la BNF est de collecter et préserver une sélection de sites Internet dans le cadre du dépôt légal. Cette mission doit être menée dans le respect de la propriété intellectuelle et de la protection des données personnelles, ce qui rend les collections non accessibles en ligne, et fait du Web une composante parmi d'autres d'un patrimoine éditorial français dont la mémoire doit être préservée. Ce panorama offre un contraste net avec la mission que s'est assignée l'Archive Team, précédemment évoquée, qui n'est contrainte « que » par la disponibilité des ressources informatiques et le souhait, de la part des utilisateurs, de les partager. Dans le premier cas, on voit le poids d'un héritage historique et de questions de souveraineté liées au dépôt légal et dans le second, le lien entre la capacité technique de l'individu et sa possibilité de contribuer à l'entreprise d'archivage.

L'archivage du Web révèle également la présence de tensions géopolitiques, illustrées de façon emblématique par les appels de Brewster Kahle lors du blocage d'Internet Archive par la Chine (Kahle 2014) ou lors des élections présidentielles américaines de novembre 2016, lorsqu'il appelle, à la suite de la victoire de Donald Trump, à un financement participatif pour créer par précaution une copie complète des collections numériques de l'Internet Archive (Kahle 2016).

On retrouve aussi dans l'archivage du Web certaines dynamiques qui rappellent le problème de la fracture numérique : cette communauté inclut presque exclusivement des institutions du « Nord global » (Gomes, Miranda et Costa 2011) – la présence des pays en voie de développement dans le Web archivé n'étant aucunement proportionnelle à leur présence croissante au sein du Web vivant. Un certain nombre d'associations régionales pourrait épauler l'action globale de l'IIPC et faire office de « sous-forums » pour l'échange

20 Voir notamment (Paloque-Bergès 2017).

21 Cf. https://web90.hypotheses.org/.

22 Cf. http://resaw.eu/.

23 La citation originale est la suivante : « *We thought the machines were going to save us – crawling the web, digitizing the books, organizing the information – but we were wrong [...] Communities of people are at the heart of curation* ».

entre acteurs autour de problèmes spécifiques à certaines régions et 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* » (Duranti et Shaffer 2012), parmi les problèmes élémentaires de l'archivage numérique se trouvent la simple absence de ressources techniques, légale et financière, comme dans le cas de la sauvegarde des archives juridiques du Burundi. Pour pallier le risque de perdre des ressources culturelles, politiques et sociales importantes, certaines institutions « du Nord » ont entrepris d'en préserver certaines (par exemple, l'université d'Heidelberg effectue une collecte du Web socio-politique chinois) ; mais à long terme, une réponse durable devra sans doute résider dans le développement d'initiatives locales.

Enfin, on retrouve dans l'archivage du Web la dialectique entre différentes pratiques et sources de normativité, de la technologie au marché, de la concertation transnationale et internationale aux standards et aux droits – une pluralité d'instruments de gouvernance qui avait déjà été identifiée pour la gouvernance d'Internet (Bygrave et Bing 2009 ; Badouard *et al.* 2013). Le « sauvetage » de Geocities opéré par l'Internet Archive 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 « standardisante » de l'IIPC : ces différents instruments de gouvernance co-existent 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 (« Report of the Working Group on Internet Governance » 2005).

Bibliographie

Ankerson, Megan Sapnar. 2015. « Take Me Back ! Web History as Chronotourism of the Digital Archive ». Communication présenté à Times and Temporalities of the Web, Paris, France.

Atkinson, Sarah et Sarah Whatley. 2015. « Digital Archives and Open Archival Practices ». *Convergence* 21 (1) : 3-7. https://doi.org/10.1177/1354856514560292.

Badouard, Romain. 2016. « “Je ne suis pas Charlie”. Pluralité des prises de parole sur le web et les réseaux sociaux ». Dans *Le Défi Charlie. Les médias à l’épreuve des attentats*, par Pierre Lefébure et Claire Sécail. Lemieux Éditeur. https://hal.archives-ouvertes.fr/hal-01251253.

Badouard, Romain, Francesca Musiani, Cécile Méadel et Laurence Monnoyer-Smith. 2013. « Towards a Typology of Internet Governance Socio-technical Arrangements ». Dans *Normative Experience in Internet Politics*, par Françoise Massit-Folléa, Cécile Méadel et Laurence Monnoyer-Smith, 99-124. Paris, France : Presses des Mines ; OpenEdition.

Barats, Christine, éd. 2013. *Manuel d’analyse du web en Sciences Humaines et Sociales*. Armand Colin. https://doi.org/10.3917/arco.barat.2013.01.

Ben-David, Anat et Adam Amram. 2018. « The Internet Archive and the Socio-technical Construction of Historical Facts ». *Internet Histories* 2 (1-2) : 179-201. https://doi.org/10.1080/24701475.2018.1455412.

Boullier, Dominique. 2015. « Charlie est un phénomène de 3eme génération (aussi) par D. Boullier ». Billet. *SHS 3G* (blog). 1 juin 2015. https://shs3g.hypotheses.org/114.

Brügger, Niels. 2009. « Website History and the Website as an Object of Study ». *New Media & Society* 11 (1-2) : 115-132. https://doi.org/10.1177/1461444808099574.

—. 2011. « Web Archiving : between Past, Present, and Future ». Dans *The Handbook of Internet Studies*, édité par Mia Consalvo et Charles Ess, 24-42. Oxford, Royaume-Uni : Wiley-Blackwell. https://doi.org/10.1002/9781444314861.ch2.

—. 2012. « Web History and the Web as a Historical Source ». *Zeithistorische Forschungen. Studies in Contemporary History* 9 (2) : 316-325. https://doi.org/10.14765/zzf.dok-1588.

Bygrave, Lee A et Jon Bing. 2009. *Internet Governance : Infrastructure and Institutions.* Oxford, Royaume-Uni : Oxford University Press. http://www.SLQ.eblib.com.au/patron/FullRecord.aspx?p=430398.

Chabin, Marie-Anne. 2012. « L’ère numérique du faux ». *Médium* 31 (2). https://doi.org/10.3917/mediu.031.0046.

Clavert, Frédéric. 2017. « Le goût de l'API ». *Le Goût de l'archive à l'ère numérique* (blog). 20 octobre 2017. http://www.gout-numerique.net/table-of-contents/gout-api.

DeNardis, Laura. 2014. *The Global War for Internet Governance*. New Haven, États-Unis : Yale University Press. https://doi.org/10.12987/yale/9780300181357.001.0001.

Duranti, Luciana et Elizabeth Shaffer, éd. 2012. « The Memory of the World in the Digital Age : Digitization and Preservation. An international Conference on Permanent Access to Digital Documentary Heritage ». Dans *Conference Memory of the World 20th Anniversary*. Vancouver, Canada : UNESCO.

Finnemann, Niels Ole. 2015. « Hypertextual Relations in Digital Born Materials : Hypertext and Time : Towards a Genre Analysis of Heterogeneous Digital Materials ». Dans *Web Archives as Scholarly Sources : Issues, Practices, Perspectives*. Aarhus, Danemark.

Giglietto, Fabio et Yenn Lee. 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 ». Dans *#Microposts2015*, 1395 : 33-37. http://ceur-ws.org/Vol-1395/.

Gomes, Daniel, João Miranda et Miguel Costa. 2011. « A Survey on Web Archiving Initiatives ». Dans *Research and Advanced Technology for Digital Libraries*, édité par Stefan Gradmann, Francesca Borri, Carlo Meghini et Heiko Schuldt, 6966 : 408-420. Berlin, Allemagne : Springer. https://doi.org/10.1007/978-3-642-24469-8_41.

Kahle, Brewster. 2014. « Please Help Protect Net Neutrality ». *Internet Archive Blogs* (blog). 10 septembre 2014. https://blog.archive.org/2014/09/10/please-help-protect-net-neutrality/.

—. 2016. « Help Us Keep the Archive Free, Accessible, and Reader Private ». *Internet Archive Blogs* (blog). 29 novembre 2016. https://blog.archive.org/2016/11/29/help-us-keep-the-archive-free-accessible-and-private/.

Kessous, Emmanuel. 2012. *L'Attention au monde : sociologie des données personnelles à l'ère numérique*. Paris, France : Armand Colin.

Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine et Rachel Donahue. 2010. *Digital Forensics and Born-digital Content in Cultural Heritage Collections*. Washington, États-Unis : Council on Library and Information Resources.

Merzeau, Louise. 2012. « Faire mémoire de nos traces numériques ». *E-dossier de l'audiovisuel*, juin. https://halshs.archives-ouvertes.fr/halshs-00727308/.

—. 2014. « Vers un Web temporel ? Constituer des corpus pour la recherche contemporaine : de l'archivage du Web à son analyse ». Communication présenté à Conférence du consortium international pour la préservation de l'internet (IIPC), Paris, France.

Musiani, Francesca et Valérie Schafer. 2019. « Science and Technology Studies Approaches to Web History ». Dans *The SAGE Handbook of Web History*, par Niels

Brügger et Ian Milligan, 73-85. Londres, Royaume-Uni : SAGE. https://halshs.archives-ouvertes.fr/halshs-02320717.

Niu, Jinfang. 2012. « An Overview of Web Archiving ». *D-Lib Magazine* 18 (3/4). https://doi.org/10.1045/march2012-niu1.

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 ». Dans *Heritage and Digital Humanities : How should training practices evolve?*, par Bernadette Nadia Saou-Dufrêne et Benjamin Barbier, 277-290. Zurich, Suisse : LIT.

—. 2017. « Usenet as a Web Archive. Multi-layered Archives of Computer-mediated Communication ». Dans *Web 25 : Histories from the First 25 Years of the World Wide Web*, par Niels Brügger, 112 : 227-250. New York, États-Unis : Peter Lang.

—. 2018. *Qu'est-ce qu'un forum internet ? : Une généalogie historique au prisme des cultures savantes numériques*. OpenEdition Press. https://doi.org/10.4000/books.oep.1843.

« Report of the Working Group on Internet Governance ». 2005. Bogis-Bossey, Suisse : Working Group on Internet Governance. https://www.wgig.org/docs/WGIGREPORT.pdf.

Schafer, Valérie. 2015. « En construction : la fabrique française d'Internet et du Web dans les années 1990 ». HDR, Paris, France : Université Paris-Sorbonne.

Schafer, Valérie, Francesca Musiani et Marguerite Borelli. 2016. « Negotiating the Web of the Past : Web Archiving, Governance and STS ». *French Journal For Media Research* 6. http://frenchjournalformediaresearch.com/lodel/index.php?id=952.

Sire, Guillaume. 2015. « Inclusion exclue : le code est un contrat léonin: Enquête sur la valeur technique et juridique du protocole robots.txt ». *Réseaux* 189 (1). https://doi.org/10.3917/res.189.0187.

Smit, Eefke, Jeffrey Van Der Hoeven et David Giaretta. 2011. « Avoiding a Digital Dark Age for Data : Why Publishers Should Care about Digital Preservation ». *Learned Publishing* 24 (1) : 35-49. https://doi.org/10.1087/20110107.

Winner, Langdon. 1980. « Do Artifacts Have Politics? » *Daedalus* 109 (1) : 121-136.