Francesca Musiani, Camille Paloque-Bergès, Valérie Schafer et Benjamin G. Thierry

Qu'est-ce qu'une archive du web ?

OpenEdition Press

Où commence et s'arrête l'archive ?

DOI : 10.4000/books.oep.8743
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2019
Collection : Encyclopédie numérique
ISBN électronique : 9791036504709

OpenEdition books

http://books.openedition.org

Référence électronique
MUSIANI, Francesca ; et al. *Où commence et s'arrête l'archive ?* In : *Qu'est-ce qu'une archive du web ?* [en ligne]. Marseille : OpenEdition Press, 2019 (généré le 17 mars 2020). Disponible sur Internet : <http://books.openedition.org/oep/8743>. ISBN : 9791036504709. DOI : https://doi.org/10.4000/books.oep.8743.

View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by OpenEdition

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).