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

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

OpenEdition Press

Des archives comme les autres ?

DOI : 10.4000/books.oep.8740
Éditeur : OpenEdition Press
Lieu d'édition : OpenEdition Press
Année d'édition : 2019
Collection : Encyclopédie numérique
ISBN électronique : 9791036504709

DES ARCHIVES COMME LES AUTRES ?

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

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

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

Une brève histoire de l'archivage du Web

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Le cas européen

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

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

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

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

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

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

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

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

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

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

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

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

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

Une composante du patrimoine nativement numérique

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

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

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

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

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

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

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

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

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

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

Les archives du Web entre rupture et continuité

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

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

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

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

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

Ces éléments impactent les métiers des archives comme des bibliothèques :

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

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

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

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

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

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

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

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

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

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

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