Archives

ARCHIVES

L'archivage du Web : bibliothèques et archives à la croisée des chemins

Clément Oury

Volume 47, numéro 1, 2017

URI : https://id.erudit.org/iderudit/1041828ar
DOI : https://doi.org/10.7202/1041828ar

Aller au sommaire du numéro

Éditeur(s)

Association des archivistes du Québec (AAQ)

ISSN

0044-9423 (imprimé)
2369-9256 (numérique)

Découvrir la revue

Citer cet article

Oury, C. (2017). L'archivage du Web : bibliothèques et archives à la croisée des chemins. *Archives*, *47*(1), 107–124. https://doi.org/10.7202/1041828ar

Résumé de l'article

Le Web est devenu, en quelques années, un support essentiel de diffusion de l'information. Cependant, la nature même du Web remet en cause une des fonctions principales des institutions culturelles : la conservation du patrimoine. Cela est d'autant plus le cas pour les bibliothèques nationales qu'elles sont souvent en charge, au titre de leur mission de dépôt légal, de la collecte et de la conservation de l'ensemble de la production scientifique et culturelle d'une nation. Pour faire face aux défis que représentait l'archivage d'une masse de données aussi vaste et aussi hétérogène, les bibliothèques ont donc été amenées à questionner leurs approches. Elles ont dû notamment s'inspirer des démarches d'autres communautés professionnelles, particulièrement celle des archivistes. La terminologie et certaines méthodes archivistiques ont souvent été employées : accent mis sur la collecte, pratique de l'échantillonnage… De fait, les collections constituées ont elles-mêmes un statut hybride, tenant à la fois de la publication et du document d'archives. Puisque les archives du Web sont des artefacts, des agrégats complexes, la question de l'authenticité prend également une place déterminante. L'ouverture à de nouvelles approches– qui ne signifie pas forcément une fusion des identités professionnelles – doit inciter à la coopération entre institutions.

ÉTUDE

L'archivage du Web : bibliothèques et archives à la croisée des chemins

CLÉMENT OURY
Chef du service Données, Réseau et Normes, Centre ISSN International

INTRODUCTION

Combien de temps consacrons-nous à utiliser les services de l'Internet ? Quelle part de notre activité, quelle part de notre vie est passée devant des écrans – d'ordinateur, de tablette et, de plus en plus, de téléphone ? Diverses études ont démontré le rôle considérable que jouent désormais les contenus numériques diffusés en ligne, dans nos pratiques professionnelles comme de loisirs[1]. Quel média a déjà pris, dans un laps de temps aussi court, une importance aussi déterminante ? De fait, si les origines de l'Internet remontent aux années 1960, le Web, né au tournant des années 1990, n'a guère plus de vingt-cinq ans.

Les bibliothèques et les autres institutions culturelles ont appris à utiliser les réseaux numériques pour perpétuer leur mission de diffusion des connaissances, que l'on songe à des activités aussi différentes que les services de référence en ligne ou la mise en place de bibliothèques numériques. En revanche, la nature même du Web les a remises en cause dans une de leurs fonctions essentielles : la conservation du patrimoine.

Cela a été d'autant plus le cas pour les bibliothèques nationales puisqu'elles sont souvent chargées, au travers de leur mission de dépôt légal, de la collecte et de la conservation de l'ensemble de la production scientifique et culturelle d'une nation. La masse de données à prendre en compte, les caractères hétérogène et incontrôlé des contenus ainsi que la structure réticulaire des sites qui les diffusent ne les ont pas seulement amenées à employer de nouvelles technologies d'archivage Web, elles les ont surtout contraintes à remettre en question leurs approches et à s'inspirer des démarches d'autres communautés professionnelles, particulièrement celle des archivistes.

L'objet de cet article est de s'interroger sur une possible hybridation des pratiques professionnelles devant les défis de l'archivage du Web. On s'intéressera à l'influence de la terminologie et des pratiques archivistiques dans le domaine de la constitution des collections. On se demandera aussi si ces archives du Web sont plus proches des types de ressources conservées par les bibliothèques ou les services d'archives. Enfin, en se posant la question de l'authenticité des archives du Web, on sera amené à voir comment des principes tirés de la description archivistique peuvent influencer le choix des métadonnées à conserver.

1. LA FRAGILITÉ DES CONTENUS WEB

Cependant, pour comprendre les problèmes que pose le Web en termes de conservation aux institutions patrimoniales, il faut revenir aux sources de son succès. Divers facteurs ont favorisé le développement sans précédent d'un média tourné vers l'innovation. On en retiendra deux qui bouleversent les approches traditionnelles des bibliothèques à la fois des points de vue juridique, technique et scientifique. Évoquons tout d'abord la prolifération des contenus, notamment des « contenus créés par l'utilisateur » (*user generated content*). Contrairement à tous les médias qui l'ont précédé, le Web permet une production de contenu à coût très limité, en utilisant diverses plateformes gratuites d'hébergement et de diffusion (de blogues, d'images, de vidéos, etc.). À ce titre, le Web est le premier média qui peut faire disparaître le filtre de l'éditeur : auparavant, le besoin de financement d'un contenu entraînait le recours à un tiers, qui validait la qualité – commerciale et/ou scientifique – d'un texte, d'une musique ou d'un film, avant de le produire et de le diffuser. Ce véritable changement de paradigme a permis une multiplication des contenus

disponibles en ligne même si, en termes de succès, les documents les plus consultés bénéficient souvent d'une promotion sur d'autres médias.

Le second facteur sur lequel il s'agit d'insister est ce que l'on peut appeler l'apparente ubiquité du Web : un document mis en ligne est accessible de partout, de n'importe quelle interface reliée au réseau. C'est la promesse d'un accès le plus large possible aux contenus. On peut identifier dans ces deux caractéristiques une apparente démocratisation : démocratisation de la production par la suppression du filtre éditorial et démocratisation de la consultation par l'accès universel aux contenus – du moins à ceux qui sont librement accessibles. Cependant, les facilités d'édition et de diffusion offertes par les technologies numériques présentent un revers que les organisations comme les individus expérimentent de plus en plus fréquemment : la fragilité des contenus et leur absence de pérennité en ligne.

Les risques qui pèsent sur des documents numériques en ligne sont en effet nombreux. Ils sont souvent liés à leur mode de diffusion même et à ce que l'on a pu qualifier de « cardinalité paradoxale » de l'Internet. L'apparente ubiquité des documents sur le Web est en effet un leurre dangereux : un document hébergé sur un serveur unique, dès qu'il est mis en ligne, se trouve instantanément (si son accès n'est pas réservé) mis à la disposition de tous ; mais si ce seul point d'accès vient à disparaître, et sauf s'il ait été répliqué par l'éditeur initial ou par un tiers, toute possibilité de consultation est supprimée. Cette cessation d'accès peut avoir de multiples causes. Il y a bien entendu, comme pour les autres documents numériques, la possibilité d'une défaillance matérielle ou logicielle de la part du responsable du contenu ou de son hébergeur. Mais cela peut aussi être dû à une altération de l'architecture du site : changement du nom de domaine, rachat de ce même nom par un tiers, modification de la structure des adresses URL, celles qui identifient les contenus en ligne. Enfin, le document peut avoir disparu pour des raisons qui ne sont pas propres aux technologies employées : la faillite ou la cessation d'activité d'un éditeur de contenu, le désir de l'auteur de supprimer des contenus qu'il ne juge plus pertinents ou qu'il préfère publier sur d'autres supports.

De telles disparitions de contenus ont lieu pour tous les types de producteurs et tous les types de documents : qui n'a pas expérimenté la déception de se retrouver face à une « erreur 404 », qui marque l'absence d'un contenu en ligne ? Une étude menée par l'équipe d'archivage du

Web de la British Library a montré que 40 % des fichiers collectés en 2013 avaient disparu en 2015, soit en l'espace de deux ans (Jackson 2015)[2]. L'Association française pour le nommage Internet en coopération (AFNIC), l'organisme qui gère le domaine .fr, a montré que moins de 80 % des noms de domaines en .fr sont renouvelés chaque année (AFNIC 2014)[3].

Cet état de fait est d'autant plus préoccupant que les contenus qui disparaissent représentent souvent les équivalents, sous forme numérique, de ressources qui sont collectées sur support physique par les institutions patrimoniales, notamment les bibliothèques. L'immense majorité des produits intellectuels et culturels ont en effet effectué leur transition numérique, que celle-ci ait abouti à un modèle hybride (avec coexistence des productions sur support et en ligne) ou à un modèle exclusivement en ligne. Certains domaines, comme la production scientifique ou le son, ont pu être des précurseurs, tandis que d'autres – celui du livre où le format numérique a parfois du mal à s'imposer – sont plus en retard ; mais aucun n'y échappe. Pour les institutions culturelles, l'archivage et la préservation de ces contenus deviennent donc un impératif majeur de continuité de leur mission et de leurs collections.

2. LE DÉPÔT LÉGAL DES PUBLICATIONS EN LIGNE : PRINCIPES ET LIMITES

Avant même de s'interroger sur les modalités techniques de collecte et de conservation des ressources en ligne, il importe d'en déterminer le statut. Juridiquement, les contenus librement diffusés en ligne peuvent être considérés comme des publications : ils sont en effet mis à la disposition du public. C'est ainsi que la jurisprudence française considère que sont applicables aux publications en ligne les mêmes dispositions que celles de leurs équivalents sur support : la diffamation ou encore l'incitation à la haine raciale y sont par exemple prohibées. C'est ce statut qui explique que les publications en ligne, en France comme dans de nombreux autres pays, fassent l'objet d'un dépôt légal. C'est la loi du 1er août 2006, désormais codifiée au sein du Code du patrimoine, qui a mis en place en France un dépôt légal des « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique. » (Legifrance 2009)[4] L'approche française vis-à-vis du dépôt légal de l'Internet – qui ne diffère pas, de ce point de vue, d'autres pays européens comme le Royaume-Uni ou le Danemark – est donc

résolument non sélective : ce sont tous les contenus en ligne, quelles que soient les « valeurs » scientifique, intellectuelle ou culturelle que l'on peut leur attribuer, qui peuvent faire l'objet d'une capture par les établissements dépositaires.

Une vision aussi englobante du dépôt légal n'est cependant pas partagée par tous les pays. Certains, comme les Pays-Bas, n'ont pas de tradition de dépôt légal et ont adopté une approche sélective pour leur collecte. D'autres bibliothèques, comme Bibliothèque et Archives Canada (BAC) ou Bibliothèque et Archives nationales du Québec (BAnQ), ont mis en place un système qui distingue la logique de dépôt unitaire systématique des publications numériques (de préférence aux formats PDF et EPUB) et une collecte complémentaire de sites Web. Pour ses collectes, BAnQ a besoin d'obtenir une licence, c'est-à-dire une autorisation d'archivage, auprès des ayants droit ; BAC peut requérir une « remise sur demande » des sites Web aux ayants droit (malgré ce terme juridique de remise, la technique mise en œuvre est toujours celle de la collecte)[5].

Cependant, une telle approche se trouve vite confrontée à de nombreux défis. Le premier est bien entendu celui de la masse des informations à archiver : le seul domaine .fr comporte près de trois millions de noms de domaines, et encore ! ce chiffre ne couvrirait que le tiers des sites français (AFNIC 2012)[6]. Ces millions de sites abritent des milliards de pages, régulièrement mises à jour. Cependant, l'enjeu n'est pas seulement technique, il est aussi scientifique et patrimonial. De fait, on ne trouve pas seulement, sur les réseaux numériques, les équivalents dématérialisés des publications qui existaient auparavant sur support. L'Internet et son application la plus populaire, le Web, ont entraîné la création d'objets documentaires nouveaux, qui n'ont qu'un rapport ténu avec leurs prédécesseurs sous forme physique. On peut certes arguer que les blogues représentent la version numérique des journaux personnels ou des correspondances échangées entre écrivains, qui pouvaient se voir publiées malgré leur appartenance originelle au domaine de l'intime. Mais que dire des interactions qui se passent sur les réseaux sociaux ? Pourtant, on ne peut pas se désintéresser de ces nouvelles formes de communication et de publication sous prétexte qu'elles n'auraient pas d'antécédents sur support : ainsi, on voit sur Facebook ou Twitter des écrivains discuter de leur ouvrage en cours d'écriture (Harrad 2012)[7], tandis que certains hommes politiques réservent aux réseaux sociaux la primeur des informations qu'ils diffusent (Ottenheimer 2014)[8]. De fait,

le Web n'est plus seulement un espace de publication et d'échange de documents. Il est devenu un lieu de communication, d'interactions (politiques, sociales, culturelles, commerciales, de loisirs, etc.), un lieu de vie en somme. Si on se doit de conserver la trace de cet espace mouvant qu'est le Web, comment peut-on espérer en faire un *dépôt légal* ?

On voit de fait que l'approche bibliothéconomique se trouve interrogée par le défi que représente le dépôt légal de l'Internet. D'une part, l'approche exhaustive du dépôt légal n'est plus tenable face à l'ampleur du champ à couvrir. D'autre part cependant, la constitution de collections, c'est-à-dire la sélection de contenus en fonction du seul critère de leurs qualités scientifique, intellectuelle ou artistique, ou de leur lien à un sujet ou un territoire donné, risque de négliger une importante part du patrimoine numérique, celui-là même qui est le plus innovant et se rattache le moins à la production éditoriale sur support – sans compter que ce patient travail de sélection demanderait des ressources humaines considérables. Il est dès lors intéressant de se tourner vers les démarches développées par d'autres professionnels du patrimoine, notamment les archivistes.

3. L'INFLUENCE DES PRATIQUES ARCHIVISTIQUES DANS LA CONSTITUTION DES COLLECTIONS

Historiquement, ce sont plutôt les bibliothèques qui se sont lancées les premières dans l'archivage du Web : en 1996, de tels projets ont démarré de façon concomitante dans les bibliothèques nationales de Suède et d'Australie, tandis que la fondation américaine Internet Archive, qui se définit elle-même comme une « *digital library* », entreprenait de moissonner le Web mondial. Si l'on regarde le paysage des institutions patrimoniales aujourd'hui, c'est là encore les bibliothèques qui dominent : le consortium international pour la préservation de l'Internet (IIPC) comporte une cinquantaine de membres, dont une trentaine de bibliothèques nationales et une dizaine de bibliothèques universitaires, contre seulement trois archives nationales[9]. En effet, alors que les bibliothèques commençaient à s'intéresser aux publications en ligne – notamment celles du Web –, les archivistes se concentraient davantage sur les défis de la production documentaire administrative sous forme électronique.

Et pourtant, les pratiques qui ont été développées dans de nombreux pays ressemblent souvent davantage à celles du monde des archives. Cela se manifeste tout d'abord dans le vocabulaire. Il ne faut pas trop focaliser sur le terme « archivage » du Web souvent employé : il n'est que la transposition, en français, de *Web archiving* utilisé de longue date par les anglophones. D'autre part, ce terme repose sur une définition très large de la notion « d'archivage » : si l'on en croit l'encyclopédie en ligne Wikipédia – qui a le mérite de fournir des définitions communément acceptées – l'archivage électronique est :

> …l'ensemble des actions, outils et méthodes mis en œuvre pour réunir, identifier, sélectionner, classer et conserver des contenus électroniques, sur un support sécurisé, dans le but de les exploiter et de les rendre accessibles dans le temps. (Wikipédia 2016)[10]

Cela couvre d'évidence le périmètre d'activité de nombreux professionnels de l'information, et n'est pas limité au seul domaine d'expertise des archivistes.

En revanche, d'autres termes sont plus significatifs, car ils renvoient à des pratiques particulières. En lieu et place du dépôt (où les éditeurs envoient leur production à la bibliothèque) ou de l'acquisition à titre onéreux (où le bibliothécaire achète le titre qu'il a choisi), l'archivage du Web repose sur la « collecte », terme qui renvoie au premier des quatre « C » de l'archiviste, avec le classement, la conservation et la communication. Il s'agit d'aller chercher les contenus là où ils se trouvent[11]. À cette fin, la plupart des institutions utilisent des robots, ou « collecteurs » (le terme anglais *crawler* est souvent utilisé) : ce sont en fait des logiciels auxquels on donne une liste d'adresses URL de départ, qui se connectent aux pages correspondant à l'URL, l'archivent, l'analysent pour trouver d'autres liens et suivent ces liens pour découvrir encore d'autres contenus. Ainsi, les robots utilisent la trame hypertextuelle du Web pour réaliser leur mission patrimoniale. En fonction des instructions qui leur sont données, les robots décideront de collecter en profondeur (en privilégiant, voire en se cantonnant aux liens à l'intérieur du ou des sites donnés) ou en largeur, en tentant de suivre le plus grand nombre de liens sortants.

La ressemblance avec l'archivistique ne se cantonne toutefois pas aux méthodes employées pour constituer les collections. Pour marier

l'inconciliable – le dépôt exhaustif et la sélection documentaire – les institutions responsables du « dépôt légal » de l'Internet ont développé des approches originales qui reposent sur l'échantillonnage. C'est une attitude que les archivistes ont expérimentée depuis longtemps pour faire face à la gigantesque masse documentaire qu'ils devaient traiter, tout en conservant une image que l'on pouvait espérer représentative de l'ensemble (Délégation interministérielle aux Archives de France 2014)[12]. Prenons l'exemple de la Bibliothèque nationale de France (BnF) qui a développé un modèle intégré pour son dépôt légal de l'Internet. Celui-ci associe, d'une part, des collectes « larges » de l'ensemble des sites en .fr et au-delà. Cette capture permet d'enregistrer des millions de sites, soit plus d'un milliard de fichiers par an. Cependant, elle n'est parfois pas suffisamment profonde, et elle ne permet pas d'archiver de façon complète les sites les plus volumineux, ceux qui représentent individuellement plusieurs milliers de fichiers. En outre, elle n'est réalisée qu'une fois par an, ce qui ne permet pas de conserver les états intermédiaires des sites. Ainsi, à défaut de pouvoir tout prendre, on essaie de prendre un peu de tout, de faire un échantillonnage.

C'est pourquoi, d'autre part, la BnF collecte à une profondeur supérieure et à des fréquences plus régulières un ensemble de sites sélectionnés par ses agents ou ceux d'institutions partenaires – grandes bibliothèques en région, bibliothèques universitaires ou laboratoires de recherche. Ceux-ci peuvent être choisis en fonction de leur sujet, de leur taille, de leur fréquence de mise à jour (comme les titres de presse) ou de leur lien à un événement (élection, festival, compétition sportive, etc.) (Bonnel et Oury 2014)[13]. Cela permet de garantir la qualité de la capture des sites qui sont considérés comme prioritaires, soit parce qu'ils représentent les équivalents de ce que la BnF et ses partenaires recevaient auparavant sur support physique, soit parce qu'ils ressortent des formes les plus innovantes de la communication en ligne. La plupart des bibliothèques nationales membres d'IIPC ont adopté une attitude similaire, consistant à combiner échantillonnage général d'un ensemble documentaire et des collectes plus poussées de ressources jugées prioritaires.

4. LES ARCHIVES DU WEB : UN TYPE DOCUMENTAIRE COMPLEXE

Les collections constituées par ce biais peuvent-elles pour autant être considérées comme des archives ? On a vu que la plupart des lois nationales considèrent les ressources diffusées en ligne comme des publications. Cependant, si l'on en croit la définition des archives fournies dans le Code du patrimoine en France, on serait tenté d'y intégrer les collections du dépôt légal de l'Internet :

> L'ensemble des documents, quels que soient leur date, leur forme et leur support matériel, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité.[14] (Legifrance 2009)

Les captures des sites de l'État et des collectivités territoriales pourraient, au prisme de cette définition très large, être considérées comme des archives publiques – après tout, la communication d'un chef d'État ou de gouvernement sur un site officiel n'engage-t-elle pas la parole de l'État ? On trouverait de même dans les collections du dépôt légal de l'Internet des archives d'entreprises ou d'associations, tandis que des captures de sites personnels deviendraient, en tant que trace de l'activité d'un individu, des archives personnelles.

De fait, les usages que l'on a constatés, ou que l'on peut escompter, des « archives » du Web, ne sont souvent guère différents de ceux que l'on fait des archives au sens propre, c'est-à-dire les documents (d'origine publique ou privée) conservés par les services d'archives. Les archives ont une valeur patrimoniale ; elles ont vocation à servir de sources à ceux qui demain feront l'histoire de nos sociétés. Comment une histoire de la société française – ou canadienne – du premier quart du XXIe siècle, qui ne disposerait d'aucune trace de ce qui se diffusait en ligne, ne serait pas tronquée ? C'est du reste sur cette vocation historique ou patrimoniale que les missions des archives et des bibliothèques sont les plus similaires.

Mais les archives du Web peuvent aussi avoir une valeur probante. Une jurisprudence, certes peu développée, existe déjà au sujet des archives du Web constituées par Internet Archive (Soubeyrand 2014)[15]. Tout document enregistré par le robot est horodaté, c'est-à-dire que le

moment de capture est strictement enregistré. Dès lors qu'il est conservé par un tiers de confiance (comme une institution patrimoniale), il pourrait servir de preuve afin, par exemple, de démontrer la présence en ligne d'un document à une certaine date, dans le cas de litiges sur la propriété intellectuelle, ou de donner le contenu précis d'une ressource susceptible de changer. Il est déjà arrivé que certains chercheurs viennent chercher, dans les collections du dépôt légal de l'Internet de la BnF, les traces des « conditions générales d'utilisation » d'un service sur un site donné à une certaine date.

Enfin, tout comme les archives personnelles, les archives du Web peuvent jouer le rôle d'une mémoire à la fois collective et personnelle. On viendra peut-être y consulter, dans un proche avenir, un « Web perdu » empreint de nostalgie ; de même que le grand public découvre aujourd'hui les traces d'un monde passé lors des opérations de valorisation réalisées par les services d'archives – ou par les bibliothèques. D'un point de vue plus personnel, il est déjà arrivé, à la BnF comme ailleurs, que les auteurs ou éditeurs de sites disparus à la suite d'un incident technique utilisent les archives du Web pour retrouver leur production. À condition qu'ils aient donné la garantie qu'ils disposaient de l'ensemble des droits de propriété intellectuelle, certains ont également pu les dupliquer pour les mettre à nouveau en ligne.

5. RESTITUER LE WEB D'HIER : QUELLE AUTHENTICITÉ ?

La question de la restitution du passé – quelles qu'en soient les fins – amène à se poser la question, d'une part, des modalités de cette restitution et, d'autre part, de l'authenticité des collections. En termes de restitution, il importe que les archives du Web conservent leur caractère hypertextuel. Le robot d'indexation, lorsqu'il se promène de lien en lien, capture les fichiers au fur et à mesure qu'il les découvre. Il ne cherche pas, par exemple, à figer une page HTML, mais archive séparément la page et tous les éléments (images, vidéos, etc.) qui la composent[16]. C'est ensuite aux outils d'indexation et de visualisation de reproduire les pages dans leurs formes de l'époque. Lorsque l'on recherche une page et que l'on saisit son URL, l'outil de visualisation va à la fois afficher la page demandée et aller chercher l'ensemble des fichiers qui doivent s'y afficher : on assiste donc à une recomposition permanente de l'archive. On peut ensuite naviguer dans le temps (en demandant la même page

ou le même fichier à une période antérieure ou postérieure), mais aussi dans l'espace : en reconstruisant la nature hypertextuelle de la collection, l'outil de visualisation offre la possibilité de naviguer dans l'archive Web comme l'internaute de l'époque naviguait sur le Web vivant.

Cependant, cette caractéristique fondamentale de la plupart des outils d'accès aux archives du Web – restituer à la fois les contenus et les liens entre contenus – présente un certain nombre de contreparties qu'il est important d'avoir à l'esprit lorsque l'on veut utiliser les archives du Web comme sources. Tout d'abord, pour accéder au contenu, il faut que les robots l'aient auparavant archivé. Or on sait que les collections sont souvent parcellaires, en raison d'obstacles techniques ou de limites économiques. Ainsi, on peut tomber sur des liens morts, ou on peut avoir des trous au sein d'une page, car l'un des fichiers qui la composaient n'a pas été capturé. Ainsi, l'archive d'une page n'est plus la restitution fidèle de la page originelle, car du contenu en est absent. Il peut aussi arriver que le robot ait su moissonner tous les fichiers d'une même page, mais que l'outil de visualisation ne sache pas en reproduire la forme[17].

Enfin, un cas plus problématique, car plus difficile à identifier, est celui de « l'incohérence temporelle »[18]. Deux cas principaux peuvent se présenter :

- L'incohérence interne à une page : l'outil de visualisation assemble, pour reconstruire une même page, des fichiers de dates différentes. Cela ne pose généralement pas de problème : si le logo d'une institution a été archivé un jour après la page d'accueil qui l'héberge, il n'a sans doute pas changé. Mais on peut aussi afficher, sur la page d'accueil d'un site de presse d'un jour donné, l'image d'actualité du jour précédent ou du lendemain. La machine « restitue » ainsi une page qui n'a jamais existé[19].
- L'incohérence externe : lorsque l'utilisateur clique, depuis une page donnée, sur une autre page, l'outil de visualisation renvoie le fichier correspondant, à la date la plus proche de la page de départ. Il peut donc arriver que, faute de mieux, l'utilisateur soit renvoyé sur l'URL qu'il désire, mais à une date très différente ; il effectuera ainsi un invisible saut dans le temps.

Cette caractéristique des archives du Web ne veut pas dire qu'elles ne sont pas fiables – elles le sont, sous réserve d'être soumises à un examen critique, comme toute autre source documentaire. Un chercheur averti

qui navigue dans les collections pourra contrôler la date d'archivage de chaque page, pour s'assurer qu'il ne fait pas un saut dans le temps trop important. De même, comme chaque fichier est individuellement horodaté, il est possible de vérifier dans quel laps de temps tous les éléments d'une même page ont été archivés, donc d'attribuer un facteur de risque d'incohérence temporelle. Ce procédé est fastidieux, mais prendrait tout son sens en cas d'utilisation de ces collections à des fins judiciaires.

6. LES ARCHIVES DU WEB, UN ARTEFACT

S'il ne doit pas susciter une méfiance excessive envers les archives du Web, ce problème d'incohérence temporelle peut en revanche amener à reconsidérer leur statut. On serait ainsi amenés à dépasser la catégorisation de cet objet documentaire en « publications » ou en « archives ». Le Web étant un espace vivant et mouvant, on ne peut pas l'archiver comme on l'entend traditionnellement, on ne peut qu'en garder des traces. De même qu'un photographe peut documenter un espace en le prenant en photo, le robot d'indexation peut réaliser des instantanés du Web qui donneront une image de ce qu'il a été à des moments précis. Davantage que des publications ou des archives, on serait donc confronté à des artefacts, produits par un acteur du champ patrimonial bien particulier : le collecteur.

L'authenticité des archives du Web ne peut donc pas se définir comme une fidélité parfaite à l'original : comment alors l'évaluer ? Comment offrir aux chercheurs les informations indispensables à son établissement ? Quelles métadonnées retenir ? Là encore, la pratique archivistique offre des réponses, notamment celle qui consiste à identifier le producteur et à documenter son activité (Association des archivistes Français 2004)[20]. Qui est le producteur ? C'est bien entendu celui qui a constitué le site Web, l'éditeur originel (administration, entreprise, artiste, universitaire, etc.), qu'une institution patrimoniale s'attachera à décrire. C'est aussi l'institution patrimoniale elle-même : ses choix, sa politique de sélection et les ressources qu'elle a pu y consacrer ont créé la collection dont le chercheur peut disposer. Le producteur est enfin, en dernier recours, le robot d'indexation lui-même : ses caractéristiques techniques, les paramètres de collecte qu'on lui a assignés (nombre maximum de fichiers par site, nombre de liens à suivre, etc.) et le journal d'événement

décrivant son activité en ligne doivent impérativement être conservés. Tous ces éléments sont qualifiés, dans le modèle de référence pour un Système ouvert d'archivage d'information (OAIS), de métadonnées de contexte et de provenance[21]. Ce sont eux qui permettront, en comprenant comment les archives ont été constituées, de rétablir l'aspect du site Web au moment où le robot s'y est promené.

7. COOPÉRER POUR MIEUX ENRICHIR SES PRATIQUES PROFESSIONNELLES

L'archivage du Web amène donc les institutions qui en ont la charge à revisiter leurs pratiques professionnelles, à s'inspirer d'autres modèles, tout en inventant des démarches spécifiques adaptées à ce média. Les bibliothèques, qui ont traditionnellement la charge de gérer des publications, particulièrement au titre du dépôt légal, se sont emparées de la mission de collecte et de conservation de leurs équivalents en ligne. Elles en sont souvent également investies par la loi, au titre du dépôt légal. Cependant, devant un type documentaire nouveau et complexe, elles ont été amenées à s'enrichir de l'expérience des services d'archives : en adoptant des pratiques dynamiques de collecte, en se reposant sur l'échantillonnage, en mettant la question de l'authenticité des collections au cœur de leur démarche et, à cette fin, en documentant un processus de production.

Cette hybridation des pratiques professionnelles a également été rendue possible par les nombreuses coopérations qui se sont tissées et qui doivent encore se développer autour de l'archivage du Web. Ce média est tellement vaste qu'une seule institution, fut-elle Internet Archive, n'est pas en mesure de l'embrasser complètement. Au niveau international, l'exemple du consortium pour la préservation de l'Internet, déjà évoqué, montre comment une cinquantaine d'institutions, sur cinq continents, peuvent travailler ensemble. Il s'agit non seulement de développer en commun des normes, des bonnes pratiques et des logiciels libres, mais aussi de fédérer de multiples initiatives d'archivage pour assurer une couverture du Web la plus complète possible.

À un niveau national, et pour prendre à nouveau l'exemple de la France, le dépôt légal de l'Internet est partagé entre deux institutions : l'Institut national de l'audiovisuel (INA) pour les sites de la radio et la télévision, la BnF pour tous les autres sites français. Ces deux organismes ont eux-mêmes

un réseau de coopération. Ainsi, la BnF travaille avec des laboratoires de recherche, des bibliothèques universitaires ou municipales ; l'INA comme la BnF sont en train de déployer un accès distant à leurs archives du Web depuis les enceintes des grandes bibliothèques en région[22]. Les équipes de la BnF sont enfin fréquemment en contact avec des archivistes – notamment des archivistes départementaux ou en mission dans les ministères – qui les sollicitent pour capturer des productions numériques de l'État ou des collectivités territoriales susceptibles de disparaître.

CONCLUSION

L'existence d'influences réciproques, l'hybridation de pratiques, doivent-elles pour autant faire conclure à une fusion des identités professionnelles des archivistes et des bibliothécaires, à l'ère numérique ? Rien n'est moins sûr. Même s'ils sont amenés à utiliser les mêmes technologies, même si leurs démarches convergent, archivistes et bibliothécaires ne voient pas la même chose quand ils contemplent des publications en ligne. Pour un même site, l'archiviste cherchera sans doute à identifier de façon prioritaire les documents inédits, les documents d'activité qui n'existeraient qu'en ligne. Le bibliothécaire s'attachera probablement davantage à documenter le mode de communication de l'émetteur, la façon dont il se représente et dont il construit son discours. Ainsi, pour un même document ou un même ensemble de documents, ils ne chercheront pas à archiver la même chose ou de la même façon. Le rapprochement n'est pas l'indifférenciation et c'est bien souvent en comparant ses objectifs et son activité à ceux de communautés professionnelles voisines que l'on comprend ses propres spécificités.

Clément Oury

NOTES

1. Pour la France, on peut se référer aux études sur les pratiques culturelles des Français, réalisées à fréquences régulières depuis 1973. Elles permettent de suivre les variations de l'usage des différents médias. Voir ministère de la Culture et de la Communication. (s. d.). Les pratiques culturelles des Français. Repéré le 27 mars 2016 à www.pratiquesculturelles.culture.gouv.fr/

2. JACKSON, A. (2015). Ten years of the UK Web Archive: what have we saved? [Billet de blogue]. Repéré le 27 mars 2016 à http://britishlibrary.typepad.co.uk/webarchive/2015/09/ten-years-of-the-uk-web-archive-what-have-we-saved.html

3. Repéré le 27 mars 2016 à https://www.afnic.fr/fr/ressources/publications/observatoire-du-marche-des-noms-de-domaine-en-france/edition-en-ligne-2014/facteurs-determinants-des-taux-de-renouvellement-3.html

4. Article L131-2 du Code du Patrimoine. Repéré le 27 mars 2016 à http://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006074236&idArticle=LEGIARTI000020905828

5. Pour les règles et modalités de dépôt à BAC, voir : Bibliothèque et Archives Canada. (2016). Règlement sur le dépôt légal de publications. Repéré le 29 mars 2016 à http://laws-lois.justice.gc.ca/fra/reglements/DORS-2006-337/page-1.html ; et Bibliothèque et Archives Canada. (2015). Dépôt de publications numériques ou diffusées sur Internet. Repéré le 29 mars 2016 à http://www.bac-lac.gc.ca/fra/services/depot-legal/Pages/publications-numeriques-internet.aspx. Pour BAnQ, voir Bibliothèque et Archives nationales du Québec. (s. d.). Comment effectuer le dépôt de publications numériques. Repéré le 29 mars 2016 à http://www.banq.qc.ca/services/depot_legal/depot_numeriques/index.html ; et Bibliothèque et Archives nationales du Québec. (s. d.). Collecte de sites Web. Repéré le 29 mars 2016 à http://www.banq.qc.ca/services/depot_legal/collecte_sites/index.html

6. AFNIC. (2012). 34,1 % des noms de domaine enregistrés en France sont des .fr. Repéré le 27 mars 2016 à https://www.afnic.fr/fr/ressources/publications/observatoire-du-marche-des-noms-de-domaine-en-france/edition-2012/34-1-des-noms-de-domaine-enregistres-en-france-sont-des-fr-1.html

7. HARRAD, K. (2012). Twitter – the virtual literary salon. Dans *Theguardian.com*. Repéré le 27 mars 2016 à http://www.theguardian.com/books/booksblog/2012/jan/11/twitter-virtual-literary-salon

8. OTTENHEIMER, G. (2014). Pourquoi Sarkozy a choisi Facebook pour annoncer son retour. Dans *Challenges.fr*. Repéré le 27 mars 2016 à http://www.challenges.fr/economie/20140919.CHA7966/pourquoi-sarkozy-a-choisi-facebook-pour-annoncer-son-retour.html

9. Il s'agit de *The National Archives* au Royaume-Uni, Bibliothèque et Archives Canada et Bibliothèque et Archives nationales du Québec qui ressortent tous à la fois du monde des bibliothèques et des archives. Voir Site de l'IIPC (s.d.) Repéré le 27 mars 2016 à http://netpreserve.org/about-us/members

10. Archivage électronique. Dans *Wikipedia.fr*. Repéré le 27 mars 2016 à https://fr.wikipedia.org/wiki/Archivage_%C3%A9lectronique

11. Il ne s'agit pas ici d'affirmer que l'activité de collecte serait étrangère aux pratiques des bibliothèques ; elle y est employée de longue date, que ce soit pour susciter des

dons ou pour recevoir des papiers d'écrivains. Il s'agit cependant pour les bibliothèques d'une pratique plus marginale, tandis qu'elle est au cœur du métier d'archiviste.

12. Délégation interministérielle aux Archives de France. (2014). *Cadre méthodologique pour l'évaluation, la sélection et l'échantillonnage des archives publiques*. Paris : Archives de France. Repéré le 27 mars 2016 à http://www.archivesdefrance.culture.gouv.fr/static/7742

13. Bonnel, S. et Oury, C. (2014). La sélection de sites web dans une bibliothèque nationale encyclopédique : une politique documentaire partagée pour le dépôt légal de l'Internet à la BnF. *Actes de la 80e conférence IFLA*. Repéré le 27 mars 2016 à http://production-scientifique.bnf.fr/sites/default/files/107-bonnel-fr.pdf

14. Article L131-2 du Code du Patrimoine. Repéré le 27 mars 2016 à http://www.legifrance.gouv.fr/affichCode.do?idArticle=LEGIARTI000006845559&idSectionTA=LEGISCTA000006159940&cidTexte=LEGITEXT000006074236&dateTexte=20080716

15. Soubeyrand, Q. (2014). *L'archivage du Web à valeur probante. Mémoire de master Enssib*. Villeurbanne : Enssib.

16. Il faut noter que ce comportement est celui des logiciels d'archivage et de visualisation utilisés par la BnF ainsi que la plupart des institutions patrimoniales. D'autres robots peuvent copier les pages Web en en produisant une image figée, par exemple sous format PDF : cette pratique a comme défaut de ne plus permettre de naviguer à l'intérieur du Web archivé.

17. Cela peut notamment arriver lorsque l'outil de visualisation ne sait pas interpréter la feuille de style d'une page Web ou que la feuille de style n'a pas été archivée.

18. Sur la question de l'incohérence temporelle, et plus généralement sur l'utilisation de l'archive du Web comme source, voir Brügger, N. (2005). *Archiving Websites: General Considerations and Strategies*. Aarhus: Centre for Internet Research.

19. Composer une page à partir de fichiers archivés à des dates différentes est un comportement indispensable : les robots ne peuvent pas capturer simultanément tous les fichiers, et les différents éléments d'une même page sont nécessairement moissonnés à quelques dixièmes de secondes, voire à quelques secondes d'intervalle. Le risque de l'incohérence temporelle, qui existe toujours, est en fait d'autant plus grand que les fichiers sont éloignés dans le temps.

20. Association des archivistes Français. (2004). *Abrégé d'archivistique*. Paris : AAF, p. 95.

21. Le modèle OAIS (norme ISO 14721) est le modèle conceptuel de référence pour la mise en place de systèmes d'archivage pérenne.

22. En France, comme dans de nombreux autres pays, les collections du dépôt légal de l'Internet ne sont pas librement accessibles en ligne, pour des raisons juridiques : droit de la propriété intellectuelle et protection des données personnelles. L'accès est réservé à des chercheurs accrédités dans les espaces de recherche de l'INA et de

la BnF. Cependant, il est juridiquement possible de consulter ces archives dans les salles de lecture d'établissements habilités, dans chacune des régions françaises. Le déploiement technique de ces solutions d'accès est en cours.

BIBLIOGRAPHIE

AFNIC. (2012). 34,1 % des noms de domaine enregistrés en France sont des .fr. Repéré le 27 mars 2016 à https://www.afnic.fr/fr/ressources/publications/observatoire-du-marche-des-noms-de-domaine-en-france/edition-2012/34-1-des-noms-de-domaine-enregistres-en-france-sont-des-fr-1.html

AFNIC. (2014). Facteurs déterminants des taux de renouvellement. Repéré le 27 mars 2016 à https://www.afnic.fr/fr/ressources/publications/observatoire-du-marche-des-noms-de-domaine-en-france/edition-en-ligne-2014/facteurs-determinants-des-taux-de-renouvellement-3.html

ARCHIVAGE ÉLECTRONIQUE. (s.d.). Dans *Wikipédia, l'encyclopédie libre*. Repéré le 27 mars 2016 à https://fr.wikipedia.org/wiki/Archivage_%C3%A9lectronique

ASSOCIATION DES ARCHIVISTES FRANÇAIS. (2004). *Abrégé d'archivistique*. Paris, France : AAF.

BONNEL, S. et OURY, C. (2014). La sélection de sites Web dans une bibliothèque nationale encyclopédique : une politique documentaire partagée pour le dépôt légal de l'Internet à la BnF. *Actes de la 80e conférence IFLA*. Repéré le 27 mars 2016 à http://production-scientifique.bnf.fr/sites/default/files/107-bonnel-fr.pdf

DÉLÉGATION INTERMINISTÉRIELLE AUX ARCHIVES DE FRANCE. (2014). *Cadre méthodologique pour l'évaluation, la sélection et l'échantillonnage des archives publiques*. Paris, France : Archives de France. Repéré le 27 mars 2016 à http://www.archivesdefrance.culture.gouv.fr/static/7742

HARRAD, K. (2012). Twitter – The Virtual Literary Salon. *Theguardian*. Repéré le 27 mars 2016 à http://www.theguardian.com/books/booksblog/2012/jan/11/twitter-virtual-literary-salon

JACKSON, A. (2015). Ten Years of the UK Web Archive: What Have We Saved? [Billet de blogue]. Repéré le 27 mars 2016 à http://britishlibrary.typepad.co.uk/webarchive/2015/09/ten-years-of-the-uk-web-archive-what-have-we-saved.html

LEGIFRANCE. (2009). Article L131-2 du Code du patrimoine. Repéré le 27 mars 2016 à http://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006074236&idArticle=LEGIARTI000020905828

MASANÈS, J. (2012, mars). *L'archivage du Web.* [Diaporama présenté au Collège de France]. Repéré le 27 mars 2016 à http://webdam.inria.fr/College/2803.JulienMasanes.pdf

OTTENHEIMER, G. (2014). Pourquoi Sarkozy a choisi Facebook pour annoncer son retour. *Challenges*. Repéré le 27 mars 2016 à http://www.challenges.fr/economie/20140919.CHA7966/pourquoi-sarkozy-a-choisi-facebook-pour-annoncer-son-retour.html

SOUBEYRAND, Q. (2014). L'archivage du Web à valeur probante. (Mémoire de master), École nationale supérieure des sciences de l'information et des bibliothèques (Enssib), Villeurbanne, France.