Biens Symboliques / Symbolic Goods
Revue de sciences sociales sur les arts, la culture et les idées

2 | 2018
Arpenter la vie littéraire

20 ans d'archivage du web

Une nouvelle aventure pour les bibliothécaires

20 of Web Archiving : a New Adventure for Librarians

20 años de archivo de la web : una nueva aventura para los bibliotecarios

Christine Genin

Édition électronique
URL : http://journals.openedition.org/bssg/271
DOI : 10.4000/bssg.271
ISSN : 2490-9424

Éditeur
Presses universitaires de Vincennes

Référence électronique
Christine Genin, « 20 ans d'archivage du web », *Biens Symboliques / Symbolic Goods* [En ligne], 2 | 2018, mis en ligne le 12 avril 2018, consulté le 04 mars 2021. URL : http://journals.openedition.org/bssg/271 ; DOI : https://doi.org/10.4000/bssg.271

BIENS
SYMBOLIQUES

SYMBOLIC
GOODS

PRESSES
UNIVERSITAIRES
DE VINCENNES

BIENS
SYMBOLIQUES
Revue de sciences sociales
sur les arts, la culture et les idées

A Social Science Journal
on Arts, Culture and Ideas
SYMBOLIC
GOODS

n° 2 / 2018

Arpenter la vie littéraire

Surveying Literary Life

PRESSES
UNIVERSITAIRES
DE VINCENNES

20 ans d'archivage du web

Une nouvelle aventure pour les bibliothécaires

Christine Genin

L'idée de constituer une mémoire du web est née dans les années 1990, quand la toile a commencé à prendre de l'ampleur. Le web est en effet le milieu dans lequel nous baignons, l'environnement à travers lequel nous abordons les objets et les êtres qui nous entourent, mais aussi un medium et surtout un media (Bachimont 2017). Internet est aussi très vite devenu un lieu où prennent naissance des œuvres qui n'auraient pu exister ailleurs. Il devenait dès lors nécessaire de conserver une mémoire de ce territoire dont les contenus sont en outre éphémères. En 2016, selon l'Observatoire du dépôt légal, un million de nouveaux domaines ont été créés et 650 000 ont disparu.

Les premières initiatives de sauvegarde du web ont été privées, avec notamment Internet Archive fondé en 1996 en Californie par Brewster Kahle, qui pose les premières pierres d'une bibliothèque d'Alexandrie née d'une ambition folle : sauvegarder intégralement la Toile, pour permettre aux enfants du troisième millénaire de surfer non seulement sur le web présent mais aussi dans l'Internet passé. Viennent ensuite des initiatives publiques, d'abord dans les pays scandinaves, par exemple à la Bibliothèque royale de Suède, qui dès 1997 met en œuvre des robots moissonneurs (ou *web crawler*). En France, les premières réflexions et expérimentations à la Bibliothèque nationale de France (BnF) datent de la fin des années 1990 : elles ont abouti au lancement des premières collectes en 2004 et à une loi en 2006.

Je ne faisais pas partie des équipes techniques qui, les premières, engagent la réflexion à la BnF, mais, assez vite, des bibliothécaires chargés de collections plus classiques intéressés par ces nouveaux objets sont associés aux expérimentations mises en œuvre en attendant que le législateur légifère. Nous faisons alors un test sur quelques dizaines de sites pour appréhender la charge de travail

que représenterait un dépôt site par site, comparable au dépôt légal des imprimés, avec une démarche auprès des producteurs, et nous nous livrons à des comparaisons entre les choix des algorithmes des moteurs de recherche (qui se fondent sur le *page rank*, c'est-à-dire la popularité des pages web) et ceux des bibliothécaires (qui se fondent plutôt sur leur intérêt documentaire), afin de déterminer quelle serait la meilleure façon d'archiver le web.

Le dispositif légal français

Un premier projet de loi visant à donner un cadre à cet archivage est rédigé en 2001, mais il n'est jamais arrivé jusqu'au Parlement (Péjut 2017). Il a fallu attendre cinq ans pour que la loi instaure en 2006 le dépôt légal du web. Et cinq autres années pour qu'un décret de décembre 2011 le rende effectif. Un temps mis à profit pour mener à bien la préfiguration des modes opératoires.

Il fallut d'abord, en effet, convaincre les pouvoirs publics que, si l'on souhaitait le développement de la société de l'information, il était nécessaire de se préoccuper aussi de sa mémoire. Si l'État ne le faisait pas, d'autres s'en chargeraient à sa place : les premiers sites de l'Élysée et de Matignon étaient déjà conservés aux États-Unis par une institution de droit privé. Ces arguments finirent par être entendus et il fut décidé que la loi relative au droit d'auteur et aux droits voisins dans la société de l'information, dite loi DADVSI, intégrerait l'archivage du web, entendu comme un nouvel objet patrimonial. Il fallut ensuite calmer les inquiétudes économiques : Internet était un facteur de croissance et il était exclu que le dépôt légal soit une charge financière pour ses acteurs. Cette obligation devait être indolore pour les producteurs de contenus. Mais il fallait également minimiser les coûts pour les finances publiques. C'est ainsi qu'il a été décidé de mettre en place un dépôt légal sans dépôt, fondé sur une collecte automatique effectuée à l'aide de robots moissonneurs. La troisième source d'inquiétude, juridique, concernait les conditions de consultation de cette nouvelle archive, qui devaient respecter les droits patrimoniaux[1] des auteurs, producteurs et éditeurs de contenus : les règles déjà adoptées pour l'ensemble du dépôt légal sont retenues.

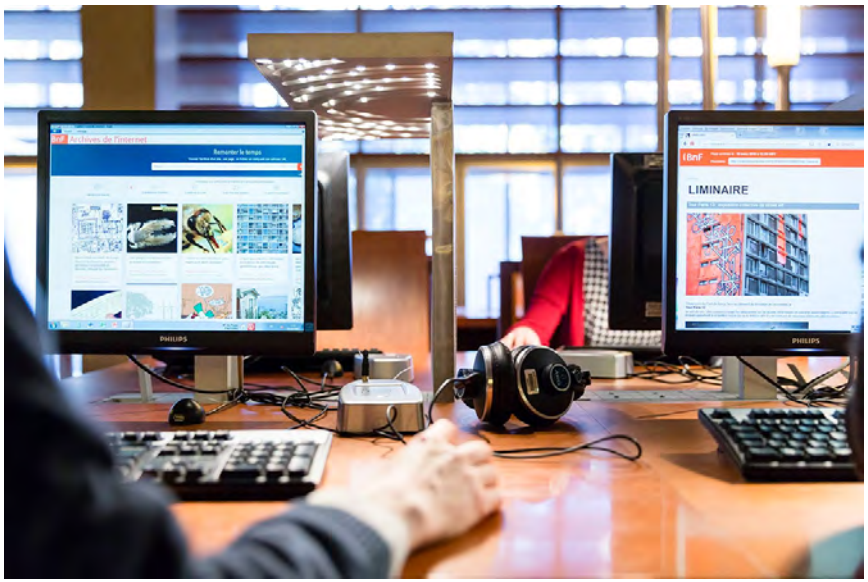

Figure 1. Consultation des Archives de l'Internet à la BnF
© BnF

La BnF est, depuis la loi DADVSI de 2006, chargée avec l'Ina (à qui revient le dépôt légal des sites de radio et télévision) de constituer une mémoire du web français. Ce dispositif a été baptisé « dépôt légal » pour marquer son inscription dans une longue histoire,

1 Voir Code du patrimoine - Titre III : Dépôt légal : partie législative ; partie réglementaire.

qui a commencé en 1537 avec l'instauration du dépôt légal des livres par François Ier, mais il s'agit plutôt d'une archive, qui ne peut que se rêver exhaustive et, à défaut, essaie au moins d'être représentative. Le dépôt légal des sites web français repose sur une définition large de l'objet patrimonial, comme « les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique », avec une limitation au web français et des possibilités de sélection et d'échantillonnage. À la différence du dépôt des imprimés, c'est un dépôt légal sans dépôt, mais qui procède par collecte automatique.

La consultation des archives se fait sur place, dans les locaux de la Bnf, et est réservée aux chercheurs accrédités. Il est assez vite apparu que cette dernière disposition était difficilement compréhensible du public – d'autant que les personnes intéressées par ces archives pouvaient déjà depuis un certain temps consulter librement en ligne les sites connectés par Internet Archive – et qu'elle serait un frein à la consultation. Un arrêté daté du 16 septembre 2014 a assoupli la règle en autorisant la consultation des collections des archives par des lecteurs accrédités, dans les 26 bibliothèques de dépôt légal imprimeur. Cette possibilité de principe est de fait proposée dans 15 établissements de province et des territoires d'outre-mer.

Dans la mesure où le web n'a pas vraiment de frontières, cette entreprise nationale s'est inscrite dès le début dans le cadre d'une coopération internationale. Le Consortium international pour la préservation de l'Internet (IIPC) a été créé en 2003 à l'initiative notamment de la BnF et compte aujourd'hui 55 institutions, réparties dans 45 pays sur tous les continents (Illien 2011). Au fur et à mesure que de nouveaux pays se dotaient d'une législation sur cette question, cette instance a permis de mettre en commun et de développer des outils en open source, de définir des normes et de réfléchir aux « bonnes pratiques ». Le projet RESAW (Research Infrastructure for the Study of Archived Web Materials) a également été créé en 2012 pour encourager une collaboration européenne.

Comment collecter le web ?

Mais constituer une mémoire du web est une entreprise complexe, à la fois sur le plan technique et sur le plan documentaire. Outre la masse des contenus, leur structuration et leur caractère dynamique constituent un défi de taille. La force du web est en effet de reposer sur une logique décentralisée où tout acteur peut devenir producteur de contenus et ce de manière désynchronisée et délocalisée. Les sites ont en outre leur propre rythme d'évolution, et leur consultation est de plus en plus contextuelle : on ne consulte jamais deux fois exactement la même page, pour paraphraser l'adage héraclitéen. Une collecte exhaustive et parfaite de la Toile est donc impossible. À défaut, il faut se donner les moyens d'un archivage le plus représentatif possible et développer des procédures permettant d'engranger des échantillons significatifs. Dans les pays ou les établissements qui se sont lancé ce défi, les stratégies sont diverses. Si la procédure de captation est forcément le plus souvent automatique, confiée à des robots qui simulent la navigation des internautes et collectent au passage les contenus visités, les modes de collecte sont des compromis : plusieurs logiques sont possibles, selon que l'on vise l'exhaustivité, un échantillonnage significatif ou une sélection, effectuée selon un algorithme de tri ou confiée à des humains.

En France, les éditeurs ou les auteurs de sites n'étant pas tenus de déposer eux-mêmes leur production, la collecte est effectuée par des robots, ou plus précisément des logiciels de capture et d'indexation. Le dispositif est sans cesse ajusté pour tenter de

suivre les évolutions très rapides du web. Il repose sur une solution mixte qui mêle deux modes d'entrée : des collectes dites larges les plus automatisées possibles et des collectes ciblées beaucoup plus fines mais limitées en taille. Au total, ce sont chaque année environ 100 téraoctets de données qui entrent ainsi dans les collections numériques de la BnF. Au-delà de ces collectes, chaque webmestre peut demander la collecte de son site.

Figure 2. Conservation des Archives de l'Internet dans SPAR
© BnF

Les collectes larges sont réalisées une fois par an ; depuis 2007 elles sont lancées à partir de la liste de tous les sites en .fr fournie par l'AFNIC (Association française pour le nommage Internet en coopération) et des listes de quelques autres hébergeurs importants, soit environ 4,8 millions de sites lors des dernières collectes. Ce mode de capture en grande partie automatisée est la meilleure façon de constituer la mémoire la plus étendue possible et de collecter le tissu d'Internet, c'est-à-dire les pages mais aussi les liens qui existent entre elles. Mais il a aussi des défauts : la profondeur de la capture est souvent insuffisante et les listes de départ risquent d'oublier les sites les moins référencés ; le domaine vaste et fuyant des blogs lui échappe par exemple assez largement.

C'est la raison pour laquelle a été ajouté un second mode d'entrée, les collectes thématiques ciblées, qui capturent de manière spécifique des sites repérés pour leur intérêt et susceptibles d'échapper à la collecte large. Ce sont des collectes beaucoup plus limitées en nombre, mais qui permettent de constituer des archives plus complètes, plus profondes et plus fréquentes. Elles sont lancées à partir de listes de sites repérés par des bibliothécaires spécialistes de chaque discipline, et parfois par des universitaires ou des partenaires extérieurs. À ce jour, près de 39 000 sites sont ainsi archivés *via* les collectes ciblées, parmi lesquels de très nombreux blogs plus ou moins personnels.

La collecte des blogs et sites personnels a d'ailleurs été, les premières années, un sujet de discussion : il a fallu lutter pour imposer aux autres bibliothécaires l'idée qu'il était également important de les collecter, au même titre que des sites institutionnels ou universitaires plus sérieux et plus faciles à assimiler aux collections déjà acquises par l'établissement. Il apparaissait d'autant plus nécessaire d'archiver ces blogs et sites personnels, qu'à la différence des sites académiques, pérennes et dotés d'archives, ils sont fragiles, éphémères, nomades, toujours susceptibles de disparaître ou de déménager.

La mise en œuvre du dépôt légal du web a ainsi conduit à une évolution du métier de bibliothécaire. La fonction de « correspondant

DL web » consiste en une veille permanente visant à ajouter de nouveaux sites ou à créer de nouveaux projets sur des thèmes émergents, mais aussi en vérifications assez chronophages destinées à éviter de continuer à collecter des sites disparus ou de perdre de vue ceux qui ont changé d'adresse. Une fois constituée cette collection d'un genre entièrement nouveau, il s'agit de l'appréhender, de s'approprier les archives explicitement collectées mais aussi celles que le robot a engrangées de lui-même, puis d'apprendre à les présenter et à les valoriser. Thibault Henneton résume à propos ces nouveaux enjeux professionnels qu'ont permis de saisir une journée d'étude organisée à l'automne 2016 à la Bibliothèque nationale de France autour de l'archivage du web :

> Les archivistes du web réunis ce 23 novembre à la BnF savent bien, vingt ans plus tard, la démesure de cette ambition. Mais ils savent également ce qu'ils ont gagné en circonscrivant leur objet. Travailler sur des corpus documentés, patiemment constitués, dans l'interdisciplinarité des sciences informatiques et sociales, c'est parvenir à une connaissance raffinée de la Toile, moins le vertige de l'exhaustivité. Ce qui s'éclaire, au terme de cette journée, ce sont les choix individuels et institutionnels qui commandent les *crawlers*, ces robots pêcheurs lancés dans l'immensité du web. C'est aussi la multiplicité des objets qu'ils ramènent dans leurs filets – dont chacun réclame un soin particulier –, les trésors qu'ils recèlent pour la science, et la diversité des métiers impliqués au cours de cette traversée. (Henneton 2017)

Archiver le web littéraire

En ce qui me concerne, je me suis particulièrement intéressée à la partie du web que l'on peut qualifier de « web littéraire ». Chargée à la BnF des acquisitions en littérature française du XXIe siècle, je constatais que de plus en plus d'écrivains créaient un site ou un blog, et que le numérique donnait une autre dimension à leur écriture. À titre personnel j'avais d'ailleurs créé en 1999 un annuaire des écrivains contemporains en ligne, puis un blog de lectures, imprimées et numériques. Avec ce que j'ai pu décrire comme « le devenir Web de la littérature » (Genin 2016), de nouvelles missions se faisaient jour pour la BnF, qui se devait d'accompagner cette évolution de la littérature. Le dispositif du dépôt légal du web lui a permis d'accomplir un travail de repérage et de collecte, en adaptant ses procédures de manière à accueillir dans ses collections ces œuvres littéraires d'un genre nouveau. À ce jour et depuis 2006, près d'un millier de sites et blogs d'écrivains français et francophones sont déjà entrés dans les collections des Archives de l'Internet en prolongement des collections imprimées.

Un autre projet dont je suis chargée est la collecte de journaux personnels en ligne (Genin 2009), lancée à partir d'une proposition de Philippe Lejeune, qui avait pris dès 2000 la mesure du passage du papier à l'écran chez les diaristes (Lejeune 2002). La coopération avec l'Association pour l'autobiographie (APA) pour la collecte des journaux personnels a permis, depuis 2007, de collecter deux fois par an plus de 1000 sites et blogs d'expression personnelle (Massip 2015). Plusieurs membres de l'APA ont apporté leur expertise dans la constitution des listes de journaux en ligne qui font l'objet de la collecte. Cette sélection, qui s'efforce de repérer les sites discrets susceptibles d'échapper aux collectes automatisées, se veut représentative des diverses tendances en matière d'expression personnelle et la plus variée possible, tant dans les contenus et les thématiques que dans les modes d'expression, l'âge et la condition sociale des auteurs. Dix ans seulement après le début de cette collecte, plus de la moitié des sites et blogs archivés ne sont plus disponibles en ligne, et avec le passage du temps leur nombre ne fera que croître, ce qui justifie d'autant plus le travail d'archivage.

De nombreuses autres collections de sites ont été constituées grâce au dépôt légal du web, couvrant toutes les disciplines traditionnelles, mais aussi les images amateurs (photos et vidéos), le web militant, la science-fiction, les sites de vulgarisation scientifique, les sites de cuisine ou les carnets de voyage en ligne. Elles apparaissent comme un prolongement naturel des collections manuscrites, visuelles, imprimées et audiovisuelles de la BnF.
L'actualité est également un axe important du dispositif. Grâce à une collecte projet qui porte ce nom, plus de 160 unes de la presse et des médias d'information, imprimés ou en ligne, sont collectées quotidiennement. Des collectes spécifiques sont aussi lancées pour conserver la mémoire des réactions à divers événements marquants, comme les attentats de janvier et novembre 2015 à Paris. Toutes les élections françaises importantes ont été couvertes très largement depuis 2005. Les Jeux olympiques donnent également lieu depuis 2010 à des collectes collaboratives dans les diverses instances internationales d'IIPC, qui sont reprises sur Internet Archive. Bruno Bachimont souligne très justement cette pluralité d'objet :

> Ce qui justifie le principe d'un archivage du web est pluriel : d'une part avoir des contenus propres au web, mais d'autre part récupérer par ces contenus des traces ou reflets du monde réel, de la société et de son histoire. Autrement dit, le web dans sa constitution est le reflet du monde, une manière de comprendre comment il fonctionne, la websphère étant désormais une composante du monde qu'elle permet d'observer et de comprendre (les réseaux sociaux permettent d'observer par exemple les tendances électorales, en complément des traditionnels sondages). Le web est donc une trace du monde, son archive est alors une trace de la trace, dans un redoublement documentaire. (Bachimont 2017)

Faire connaître cette mission

Ce dispositif d'archivage du web n'est pas encore très connu mais fait l'objet d'une politique de communication spécifique de la BnF. Un compte twitter (@DLwebBnF) a été créé en octobre 2011 pour faire connaître cette mission d'archivage et diffuser des informations sur ses actualités. Un carnet de recherche intitulé *Web Corpora* a en outre été mis en ligne en 2017 sur la plateforme Hypothèses. Il se présente ainsi : « les archives de l'Internet, collections de dépôt légal conservées à la BnF, qui forment un corpus de plusieurs milliards de documents – ou mieux encore une addition de corpus – d'une richesse insoupçonnée, et à disposition de la communauté scientifique ».

Figure 3. Annonce de la création du carnet de recherche *Web Corpora* sur le fil twitter de la BnF

L'organisation d'événements est un autre moyen de faire connaître cette mission du public. Pour la BnF et l'Ina, 2016 a été l'occasion de fêter un double anniversaire, à savoir les 10 ans de la loi sur le dépôt légal de l'Internet en France et les 20 ans des collections d'archives de l'Internet français. Une manifestation scientifique intitulée « Il était une fois dans le web, 20 ans d'archives de l'Internet en France » a été organisée en novembre 2016 à la BnF avec le concours de l'équipe Web90 et de l'Université Paris Lumière. Cet événement a reçu un assez large écho dans la presse et les médias.

La journée du 23 novembre 2016 a réuni des chercheurs issus de plusieurs disciplines et champs de la recherche avec des bibliothécaires et des professionnels du web faisant émerger une communauté scientifique française autour des archives de l'Internet. Elle a permis d'évoquer l'histoire des archives de l'Internet en France, depuis les premières communautés du web jusqu'aux outils d'accès actuels, et leur utilisation. Le dispositif a ainsi été analysé par plusieurs tables rondes qui ont réuni des acteurs du web, témoins ou déposants, des chercheurs usagers des archives ou spécialistes de domaines en relation (droit appliqué au numérique, méthodes quantitatives et cartographiques) et des praticiens de l'archivage du web. Une grande partie des interventions ont été ou seront mises en ligne sur le blog *Web corpora* et les vidéos sont disponibles sur le site de la BnF.

Ceux qui consultent les archives du web, que l'on peut appeler *archinautes*, sont encore peu nombreux mais des enquêtes permettent d'esquisser leurs profils (Illien & Chevalier 2011) : chroniqueurs à la recherche d'une information, d'un discours, d'un événement déjà lointain ; sociologues ou politistes, qui étudient l'impact du web sur la communication électorale, la formation des réseaux, des identités et des communautés numériques, ou encore la circulation du buzz et des controverses ; chercheurs en informatique, qui voient dans cette gigantesque base un terrain propice à l'invention et au calcul ; historiens du Net.art, sur les traces des premières œuvres numériques ; juristes qui se demandent si l'archive d'un site peut servir de preuve dans le règlement d'un litige ; particuliers à l'affût du blog d'un proche disparu, voire de leur propre site anéanti par une panne. Les chercheurs de toutes les disciplines vont aussi être amenés à ajuster leurs pratiques (Schafer & Thierry 2015). Pour la Bnf, il importe de faciliter l'appropriation de ces archives d'un nouveau genre, qui supposent de « développer une nouvelle herméneutique » que l'on peut rapprocher de celle des philologues, pour « exploiter des archives qui sont par essence fautives et incomplètes mais néanmoins fiables et exploitables » (Bachimont 2017).

Baliser les archives du web

La BnF conserve ainsi des fragments du web tel qu'il existe depuis plus de 20 ans. Ces strates se superposent pour former une masse déjà imposante de 29 milliards de fichiers collectés soit 794 téraoctets de données. Ces données sont d'autant moins évidentes à appréhender qu'il n'existe pas de catalogue comme pour les autres types de documents, mais seulement une application qui permet de localiser l'archive d'un site à partir de son adresse. De plus, les fonds sont lacunaires, du fait de l'impossibilité de capturer tous les sites, et des obstacles auxquels se heurtent les robots : l'accès à la plupart des bases de données et à certains fichiers leur est interdit. L'archinaute qui cherche à consulter ces données reçoit alors un message d'erreur.

Le signalement et l'indexation constituent donc des enjeux cruciaux. S'il n'existe pas de liste complète des sites accessibles dans les archives de l'Internet, les fiches thématiques de data.bnf.fr signalent les sites web sélectionnés par la BnF dans le cadre

des collectes ciblées. (par exemple en littérature française). Les listes des sites archivés dans le cadre des collectes électorales sont également disponibles sur data.gouv.

Quelques-uns des sites collectés apparaissent aussi peu à peu dans le catalogue général de la BnF, notamment lorsqu'un ISSN leur a été attribué, à l'image des revues passées du papier au numérique. Pour ce qui concerne les blogs d'écrivains, qui forment un corpus numérique prolongeant directement celui des livres imprimés, il semble important de les rendre plus facilement visibles et consultables par les lecteurs et les chercheurs, comme le sont les imprimés. Il est par exemple cohérent qu'un utilisateur puisse trouver au catalogue le *Tiers Livre* de François Bon au même titre que ses livres. Un des projets actuels vise ainsi à rendre visible dans le catalogue les notices de quelques centaines de blogs d'écrivains archivés au titre du dépôt légal du web et auxquels un ISSN a été attribué. Il n'est bien entendu pas question de cataloguer le web, mais de proposer une mise en valeur particulière de ce corpus.

Les parcours guidés offrent également quelques portes d'entrée pour découvrir les collections. Le parcours « (S)'écrire en ligne : journaux personnels et littéraires », que j'ai rédigé en 2009 en partenariat avec l'APA, propose ainsi quinze variations sur le thème du journal personnel en ligne, lequel a participé depuis 2003 (avec l'essor des blogs) à donner une ampleur nouvelle à la pratique du journal intime. Ce parcours ouvre des portes vers ce patrimoine éphémère et donne à lire et à voir des journaux et des brouillons d'écrivains, des carnets de lecture très divers, de la poésie expérimentale au polar, et une sélection de blogs, récits de vies, ordinaires ou pas, et intégrant photos, dessins ou vidéos. Il puise à cet effet parmi les archives les plus anciennes pour retracer les origines d'un genre dont l'intérêt, littéraire parfois, sociologique et psychologique toujours, est désormais avéré.

L'un des derniers parcours mis en ligne, entièrement rédigé par des chercheurs, concerne « Le Web des années 1990 – antérieur au tournant dit 2.0 », qui inspire déjà « une certaine nostalgie : celle d'un club encore fermé, impliquant une maîtrise technique pour produire des pages. Perçu comme un temps d'inventivité, c'est également le moment où foisonnent les fichiers .gif et bandeaux "En construction" et où les pages personnelles côtoient les sites vitrines d'institutions et d'entreprises. »

Les chantiers à venir

Le dispositif est sans cesse remis sur le métier de manière qu'il soit plus pertinent et se prête à des usages plus larges. Depuis 2016, les Archives de l'Internet disposent d'une nouvelle interface plus moderne, qui propose en page d'accueil un accès à une sélection de sites sous la forme de vignettes illustrées. A été ajouté un module de recherche par mot, développé en 2015 à partir de l'application Shine, utilisée par la British Library.

Depuis quelques années, les équipes de la BnF travaillent aussi avec des chercheurs pour explorer des formes innovantes d'accès aux collections du dépôt légal de l'Internet, et tenter de mettre en place des services, et des outils d'analyse du web, en proposant par exemple l'export ou la création de corpus à partir des résultats d'une recherche.

Une application expérimentale, *Labs*, a été développée pour valoriser et poursuivre deux approches qui aident à l'exploitation des collections. Ces développements s'inscrivent dans le cadre de CORPUS, programme de recherche visant à préfigurer un service de fourniture de corpus numériques à destination de la recherche. La BnF met en effet aujourd'hui à disposition plusieurs milliards de documents numériques très variés dont la masse de données

dépasse bien souvent les capacités de traitement de l'humain. Dans ce contexte d'abondance, mettre simplement à disposition les documents ne suffit plus : il est nécessaire de fournir des instruments qui permettent aux chercheurs d'aller au-delà du seul contenu. Le projet bénéficie du développement des méthodes de fouille de textes et de données, qui ouvrent de nouvelles perspectives pour questionner la place d'un texte dans un corpus ou interroger ses métadonnées et ses occurrences.

Figure 4. Archives de l'Internet Labs

Le projet « Archives de l'internet Labs » propose des statistiques et métadonnées qui prolongent une approche déjà utilisée pour cartographier le patrimoine numérisé lié à la Grande guerre (Chevallier 2017). L'équipe Web90, qui travaille sur le web des années 1990, a collaboré avec la BnF sur l'indexation en plein texte des « incunables du web français 1996-2000 », une collection acquise rétrospectivement auprès d'Internet Archive. Enfin, les chercheurs du projet ASAP (Archives sauvegarde attentats Paris) travaillent sur les collectes d'urgence effectuées par la BnF au moment des attentats de janvier et de novembre 2015. Un dispositif d'accueil spécifique a été mis en place pour permettre aux chercheurs d'explorer ces nouvelles possibilités dans le respect du cadre réglementaire : l'accès aux outils, aux données et aux métadonnées se fait sur des postes spécifiques et dans des conditions spécifiées par une convention de recherche.

Outre les nécessaires adaptations aux évolutions du web, il reste encore de nombreux chantiers à venir, dans le but, par exemple, de proposer une meilleure indexation des collectes proposées ou de travailler plus étroitement avec les chercheurs pour la création de corpus. Le web fait en tout cas aujourd'hui figure de « mémoire partagée » (Merzeau 2017) de l'humanité. Constituer une mémoire de cette mémoire et la partager avec le plus grand nombre est l'un des enjeux majeurs de notre temps.

Christine Genin
Chargée de collection en littérature française contemporaine et coordinatrice du dépôt légal du web pour le Département littérature et art (Bibliothèque nationale de France)
Site personnel : http://christinegenin.fr/blog/.

Références bibliographiques

Bachimont Bruno (2017). « L'archive du web : une nouvelle herméneutique de la trace ? ». *Web Corpora*. [En ligne] https://webcorpora.hypotheses.org/288 [consulté le 15 août 2017].

Chevallier Philippe (2017). « La Grande Guerre sur le web : un éclairage inédit ». *Carnet de la recherche à la Bibliothèque nationale de France*. [En ligne] https://bnf.hypotheses.org/1588 [consulté le 15 août 2017].

GENIN Christine (2009). « Collecter l'océan. L'archivage de l'intime en ligne ». *Bibliothèque(s)*, 47-48 : 50-53. [En ligne] http://www.enssib.fr/bibliotheque-numerique/documents/59735-47-48-intimites.pdf#page=52 [consulté le 15 août 2017].

GENIN Christine (2016). « Le devenir Web de la littérature ». *Revue de la BnF*, 52 : 152-162. [En ligne] https://www.cairn.info/revue-de-la-bibliotheque-nationale-de-france-2016-1-page-152.htm [consulté le 15 août 2017].

HENNETON Thibault (2017). « Construire l'histoire d'internet ». *Web Corpora.* [En ligne] https://webcorpora.hypotheses.org/200 [consulté le 15 août 2017].

ILLIEN Gildas (2011). « Une histoire politique de l'archivage du web : le consortium international pour la préservation de l'internet ». *Bulletin des bibliothèques de France*, 2 : 60-68. [En ligne] http://bbf.enssib.fr/consulter/bbf-2011-02-0060-012 [consulté le 15 août 2017].

ILLIEN Gildas & CHEVALLIER Philippe (2011). « Les archives de l'internet : une étude prospective sur les représentations et les attentes des utilisateurs potentiels ». *Site de la BnF*. [En ligne] http://www.bnf.fr/documents/enquete_archives_web.pdf [consulté le 15 août 2017].

LEJEUNE Philippe (2002). « Un an après ». *Autopacte*. [En ligne] http://www.autopacte.org/un_an_apr%E8s.html [consulté le 15 août 2017].

MASSIP Bernard (2015). « L'APA et l'expression personnelle en ligne ». *Association pour l'autobiographie*. [En ligne] http://autobiographie.sitapa.org/fonds/article/l-apa-et-l-expression-personnelle-en-ligne [consulté le 15 août 2017].

MERZEAU Louise (2017). « Mémoire partagée ». In Cornu-Volatron Marie, Orsi Fabienne, Rochfeld Judith (dir.). *Dictionnaire des biens communs*. Paris, Presses universitaires de France. [En ligne] http://merzeau.net/memoire-partagee/ et https://halshs.archives-ouvertes.fr/halshs-01546678/document [consulté le 15 août 2017].

PÉJUT Geneviève (2017). « Quand le web devient une archive : la construction du cadre légal », *Web Corpora*. [En ligne] https://webcorpora.hypotheses.org/283 [consulté le 15 août 2017].

SCHAFER Valérie & THIERRY Benjamin G. (2015). « L'ogre et la Toile. Le rendez-vous de l'histoire et des archives du Web ». *Socio*, 4 : 75-96. [En ligne] https://socio.revues.org/1337 [consulté le 15 août 2017].