

L'AUTEUR



Gildas ILLIEN est conservateur et chef du Service Dépôt légal numérique à la Bibliothèque nationale de France, à Paris.

l'expérience du furetage dans un grenier familial, voire... dans une bibliothèque.

Alors, garder la trace du Web aurait peut-être et malgré tout quelque chose de bon ? Une histoire tristement ordinaire et contemporaine l'illustre. Symbole de l'industrie dans l'Est de la France, l'entreprise *Vogica*, spécialiste depuis 1976 des cuisines et salles de bain faites sur mesure, a été placée en liquidation judiciaire le 8 novembre 2010. Une quarantaine de magasins ont fermé, laissant un millier de salariés sans emploi. Consulté un mois après, le site www.vogica.fr avait entièrement disparu ; à sa place, figurait un avis de 12 lignes signé du mandataire chargé de la liquidation judiciaire.

Où les anciens salariés de l'entreprise peuvent-ils retrouver la mémoire numérique de leur entreprise et, par exemple, montrer que six mois avant sa fermeture, la rubrique « offres d'emploi » du site proposait encore 28 postes ? Où les chercheurs qui, dans 10 ou 30 ans, écriront l'histoire du déclin du tissu industriel français retrouveront-ils les traces du développement, puis de la chute de cette société ? Où, enfin, les amateurs

de publicités retrouveront-ils les clips qui ont fait la popularité de l'enseigne ?

Toutes ces informations sont conservées à la BnF, dans ses « archives de l'Internet ». L'histoire du site de *Vogica* illustre la problématique fondamentale du patrimoine : ce qui semble anecdotique et sans valeur aujourd'hui peut, un jour ou l'autre, devenir une source précieuse pour l'histoire collective et intéresser nombre de nos concitoyens. Et le Web étant devenu un média de masse, où publier est désormais à la portée de chacun, le patrimoine national ne peut plus se limiter à la production des élites et des spécialistes. Pour la première fois, il est possible de l'étendre aux propos des citoyens les plus ordinaires. Le dépôt légal du Web répondrait-il au défi que représente l'invention d'une mémoire de masse ?

Le dépôt légal, né d'une ordonnance royale de François I^{er} en 1537, a permis de réunir la plus grande part des collections de la BnF. Il oblige tout éditeur, imprimeur, distributeur ou importateur à déposer des exemplaires de chaque document qu'il édite, imprime, produit, distribue ou importe sur le territoire français, à l'un des organismes dépositaires (BnF, INA, etc.). Le dépôt légal s'est étendu à tous les supports et modes de diffusion, jusqu'à la révolution numérique. Ainsi, depuis le vote de la loi de 2006, la BnF archive chaque jour, au moyen de robots moissonneurs, des centaines de sites Web. À chaque connexion, le robot d'archivage s'identifie auprès des serveurs de l'éditeur, déclarant son identité et les raisons légales de sa visite. La BnF archive les contenus récoltés dans de vastes baies de stockage – ses magasins numériques. Des copies des sites de journaux, de revues savantes, de créations artistiques, mais aussi des blogs, des jeux en ligne, des sites publicitaires ou commerciaux côtoient désormais les vénérables collections d'estampes, de cartes et de manuscrits. Et depuis 2008, ces archives sont accessibles au public.

À la différence du dépôt légal usuel, le dépôt légal des sites Web n'implique aucune démarche active de la part des éditeurs ou des webmasters. Les collectes se font automatiquement au moyen d'un robot du nom de *Heritrix*. C'est un logiciel libre, comme

Le circuit de traitement d'une archive Web

Une archive Web se constitue en trois étapes – collecte, conservation et signalement. Les défis à chaque étape sont la volumétrie, l'hétérogénéité et la volatilité des données : l'obtention d'une image fidèle du Web exige un parc de serveurs robustes et capables d'archiver un grand nombre de fichiers très variés dans un temps limité.

Les fichiers sont récoltés au moyen de robots logiciels nommés *crawlers* ou *spiders*, le plus utilisé par les bibliothèques nationales étant *Heritrix*. À partir d'une liste d'adresses URL, ces robots copient tous les contenus qu'ils rencontrent, comme le ferait un internaute surfant sur tous les liens d'un site. Les bases de données, les contenus payants ou accessibles par iden-

tifiant, etc., sont un obstacle, et la moisson se limite souvent à la surface du Web. Le paramétrage des robots et leur surveillance par des professionnels permettent de choisir la fréquence et la profondeur de capture des sites.

Les archives sont stockées dans un format (WARC) qui regroupe les fichiers et les métadonnées associées à la collecte (par exemple pour dater l'archive). La sauvegarde se fait sur disques ou sur bandes, dans des entrepôts numériques bien maintenus (copie sur plusieurs sites par exemple). Pour suivre l'évolution des formats et des logiciels, des campagnes de migration (d'un format de fichier vers un format plus récent) ou d'émulation (par la reconstitu-

tion artificielle de l'environnement initial de publication des sites) sont prévues.

Des techniques d'indexation permettent la recherche et la visualisation dans les archives. L'indexation en plein texte permet de rechercher par mots, comme avec un moteur de recherche en ligne. L'application *Wayback Machine* « remonte le temps » en cherchant, à partir de l'adresse d'un site et d'une frise chronologique cliquable, toutes les anciennes versions du site. D'autres techniques offrent aux chercheurs des traitements plus pointus (fouille de données, de liens, analyses sémantiques de corpus...) et des représentations cartographiques et dynamiques des données.

tous ceux que développe le Consortium international pour la préservation de l'Internet (IIPC), organisation internationale d'une quarantaine d'institutions culturelles qui ont fait alliance avec *Internet Archive* dès 2003 afin de promouvoir et de faciliter l'archivage du Web à l'échelle planétaire.

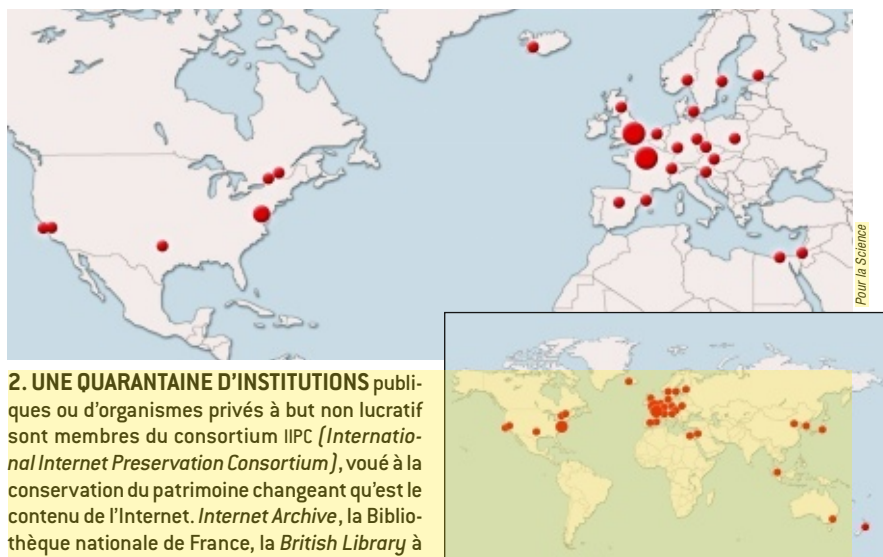
Si tout ou partie du site visité reste inaccessible au robot *Heritrix* pour des raisons techniques (base de données, contenu protégé par mot de passe, formulaire d'accès...) ou commerciales (contenu payant, abonnement...), la loi autorise la BnF à solliciter l'éditeur pour trouver des solutions techniques lui permettant d'améliorer la collecte.

Il est impossible d'être exhaustif

Compte tenu des volumes de données en ligne, la BnF ne peut garantir l'exhaustivité de son archivage. Aussi n'a-t-elle plus l'ambition ni la mission de tout conserver. Elle procède donc par échantillonnage, fondé sur la combinaison de plusieurs types de collectes (collectes automatiques et collectes ciblées, assistées par des bibliothécaires, plus fréquentes et plus profondes). La politique de la BnF est de constituer une collection représentative de l'Internet français tout en archivant certains contenus qui sont dans la continuité de son patrimoine imprimé.

Le service du « dépôt légal numérique » créé à la BnF en 2008 pour organiser et développer cette activité emploie neuf bibliothécaires et informaticiens à temps plein... ainsi qu'une quarantaine de serveurs chargés de la collecte et de l'indexation des données, une sauvegarde étant par ailleurs effectuée à la fois sur disque (pour la diffusion) et sur bande (pour la conservation). Cela représentait fin 2010 près de 200 téraoctets de données (1 téraoctet = 1 000 gigaoctets).

Ce volume peut aujourd'hui paraître modeste, mais c'est l'une des plus grandes collections d'archives Web du monde, après *Internet Archive* (près de deux pétaoctets, soit 2 000 téraoctets). Le régime juridique du dépôt légal numérique français, similaire à celui des pays nordiques, est en effet très favorable : il lui permet de collecter des sites sans avoir à demander l'accord de chaque édi-



2. UNE QUARANTAINE D'INSTITUTIONS publiques ou d'organismes privés à but non lucratif sont membres du consortium IIPC (*International Internet Preservation Consortium*), voué à la conservation du patrimoine changeant qu'est le contenu de l'Internet. *Internet Archive*, la Bibliothèque nationale de France, la *British Library* à Londres, la Bibliothèque du Congrès à Washington figurent parmi ses principaux membres. Il existe aussi d'autres acteurs de l'archivage du Web qui ne font pas partie d'IIPC, par exemple des grandes entreprises qui stockent les contenus successifs de leurs propres sites Web, notamment pour leurs besoins juridiques.

teur ou webmestre. Ce n'est pas le cas dans la plupart des pays de droit anglo-saxon. La *British Library* ou la Bibliothèque du Congrès, par exemple, ne disposent que d'archives sélectionnées et négociées à l'unité. Leurs collections couvrent certaines thématiques ou des événements (les attentats du 11 septembre 2001, l'ouragan Katrina...), mais rien qui ressemble au dépôt légal au sens où on l'entend en France.

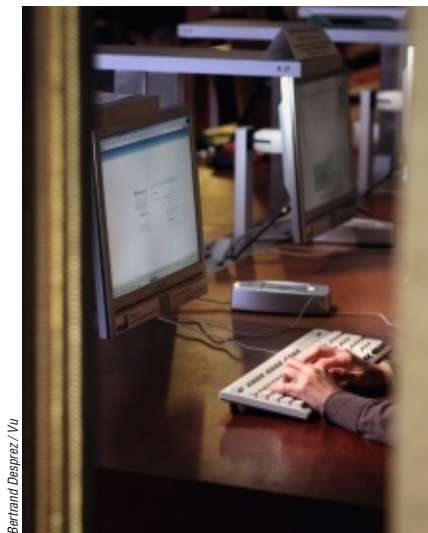
La politique patrimoniale mise en place au fil des ans à la BnF repose sur plusieurs types de collectes, visant à refléter le plus fidèlement possible la société française. Par exemple, la bibliothèque réalise une fois par an un instantané de la totalité des noms de domaines enregistrés sous « .fr » auprès de l'AFNIC (plus de 1,8 million fin 2010, qui ne représenteraient qu'un tiers des sites effectivement enregistrés en France la même année, à comparer aux 192 000 millions de noms de domaines recensés fin 2009 dans le monde). Cette collecte a occupé l'été dernier une trentaine de serveurs durant 12 semaines, et a recueilli près de 24 téraoctets de données, soit 832 millions d'URL.

Pour compléter ces collectes à grande échelle, fidèles à l'esprit du dépôt légal qui n'exclut aucun contenu *a priori*, la BnF col-

lecte aussi plus fréquemment et plus profondément des sites dont les auteurs ou les thématiques s'inscrivent dans la continuité de ses collections.

Chaque matin, le même robot va ainsi « chercher le journal », c'est-à-dire la une en ligne d'une centaine de sites d'actualités qui donnent les nouvelles du jour. Le robot veille à la pluralité des opinions et des sensibilités. Il puise aussi bien dans la presse nationale, régionale, gratuite, que dans des sites d'actualités nés avec le Web, tels *Mediapart* ou *Rue89*. Le chercheur de demain retrouvera ainsi aisément les événements marquants d'une date précise avant de plonger dans des fonds plus spécialisés qui l'attendent déjà. Ces derniers sont le fruit d'un patient travail de veille et de sélection effectué par un réseau de 80 bibliothécaires, souvent en partenariat avec des bibliothèques régionales, des associations ou des laboratoires de recherche.

Depuis 2002, chaque grande campagne électorale mobilise ainsi une équipe de la BnF qui suit jour après jour les faits et gestes numériques des candidats, de leurs opposants et partisans, et veille à leur enregistrement dans les archives portant sur les élections. On y retrouve tracts, affiches,



Bertrand Desprez / Vu

3. L'ACCÈS AUX ARCHIVES DU WEB est souvent strictement encadré en raison des données à caractère personnel qu'elles contiennent. À la BnF, cet accès est ainsi réservé aux chercheurs. En Scandinavie, les archives Web ne sont pas accessibles. En revanche, celles du site privé américain www.archive.org sont disponibles en ligne.

✓ SUR LE WEB

Le consortium IIPC :
<http://netpreserve.org/>

Dépôt légal des sites Web
à la Bibliothèque nationale
de France :

http://www.bnf.fr/fr/professionnels/depot_legal/a_dl_sites_web_mod.html

G. Illien, Le dépôt légal de l'Internet
en pratique : les moissonneurs
du Web, *Bulletin des Bibliothèques
de France*, n° 6, pp. 20-27, 2008,
disponible sur :

<http://bbf.enssib.fr/consulter/bbf-2008-06-0020-004>

Bibliographie établie par la BnF :
http://bibnum.bnf.fr/conservation/bibliographie_dl_web.pdf

CRÉDOC, La diffusion
des technologies de l'information
et de la communication
dans la société française,
décembre 2010
(rapport disponible sur
<http://www.credoc.fr/>).

discours, vidéos et autres documents numériques qui cristallisent les grands enjeux de société d'une époque. Le site de campagne du candidat Jacques Chirac archivé en 2002 (www.chiracaveclafrance.net, disparu depuis), où l'on trouve par exemple son appel aux Français le soir du premier tour de l'élection du 21 avril 2002, y fait figure de pièce de collection...

La BnF suit la vie des mouvements sociaux en ligne et conserve les sites des organisations syndicales, des grandes associations et des grandes causes, mais aussi des entreprises cotées au CAC 40 dont elle archive notamment tous les rapports annuels. Elle s'intéresse aux arts et à l'écriture littéraire. L'une de ses collections les plus fascinantes est faite de journaux personnels en ligne et de sites d'écrivains connus ou non. C'est une sorte de salon littéraire virtuel où l'on découvre les traces des premiers blogueurs [qui étaient des informaticiens] et de leurs suivants : policiers, infirmières, agriculteurs ou caissières qui, un jour, ont préféré le clavier à la plume pour raconter leur vie ou livrer leurs observations. Ces récits de vie numériques, qui côtoient les premiers brouillons et essais en ligne d'écrivains de renom, forment une galerie inédite ; pourquoi l'histoire de la littérature devrait-elle s'arrêter à Casanova, Proust ou Flaubert ?

Le robot est également programmé pour réagir aux urgences et aux grands événements imprévus. Les bibliothécaires suivent l'actualité et sont parfois confrontés à des choix difficiles. Ceux de la BnF ont, par le passé, pris des initiatives de collecte spontanée, jusque dans la rue, dont les historiens leur sont redevables aujourd'hui, comme ce fut le cas pendant l'Occupation ou Mai 68.

Ainsi, en décembre 2010, la BnF décida d'archiver l'intégralité du site controversé *Wikileaks* : l'un des serveurs miroirs étant hébergé chez OVH à Roubaix, le site était juridiquement assujéti au dépôt légal français, fondé depuis cinq siècles sur la relation au territoire national. Il ne s'agissait pas de prendre parti, mais de s'assurer que les « archinautes » du futur auraient les moyens de retrouver trace d'une affaire et d'un débat planétaires, qui seront probablement déterminants pour l'évolution de la régulation des

usages et de la liberté d'expression sur Internet. L'initiative était pertinente puisque, au même moment, le gouvernement américain interdisait l'accès au site *Wikileaks* depuis les agences fédérales américaines, y compris à la Bibliothèque du Congrès.

En prise avec l'actualité

Quelques semaines plus tard, en partenariat avec *Internet Archive*, les robots de la BnF partaient moissonner les sites de Tunisie et d'ailleurs qui ont permis de suivre les révoltes d'Afrique du Nord et du Moyen-Orient. On peut se réjouir, pour une fois, que la France ne soit pas la plus démunie juridiquement et techniquement pour se protéger, face à ce type d'événements, contre l'oubli.

Entre droit à l'oubli et devoir de mémoire numérique, on voit ainsi s'opposer des arguments convaincants de part et d'autre. Idéalement, un archivage qui ne conserverait du Web que sa partie publique, filtrant toute donnée personnelle, constituerait évidemment le compromis le plus acceptable entre la protection des individus et la conservation d'un patrimoine commun. Ce scénario est toutefois impossible à mettre en œuvre dans l'état actuel du droit et des technologies : même si les collectes de la BnF ne représentent qu'une part infime de la totalité des données publiées en ligne, pour rester représentatives elles doivent être effectuées à une échelle suffisamment grande, ce qui implique une large part d'automatisation.

Or identifier et filtrer les données personnelles à cette échelle n'est pas possible, pour des raisons économiques et informatiques, mais aussi parce qu'il est difficile d'apprécier le caractère personnel des informations. Les pratiques d'expression en ligne ont rendu très poreuses les lignes de partage habituelles entre espace public et privé, entre publication et communication interpersonnelle. De ce point de vue, les robots d'archivage ne sauraient faire la part des choses, même s'ils respectent les espaces privés lorsqu'ils sont clairement signalés comme tels (par exemple, le robot n'a pas d'amis sur *Facebook* et ne peut s'introduire dans un réseau social à moins qu'il ne soit déclaré public et visible par tous).