LE DÉPÔT LÉGAL DE L'INTERNET FRANÇAIS À LA BIBLIOTHÈQUE NATIONALE DE FRANCE

Évelyne Cohen, Julie Verlaine

Publications de la Sorbonne | « Sociétés & Représentations »

2013/1 n° 35 | pages 209 à 218

ISSN 1262-2966
ISBN 9782859447458

Article disponible en ligne à l'adresse :

---

http://www.cairn.info/revue-societes-et-representations-2013-1-page-209.htm

---

Pour citer cet article :

Évelyne Cohen, Julie Verlaine, « Le dépôt légal de l'internet français à la Bibliothèque nationale de France », *Sociétés & Représentations* 2013/1 (n° 35), p. 209-218.
DOI 10.3917/sr.035.0209

LIEUX ET RESSOURCES

Évelyne Cohen et Julie Verlaine

Le dépôt légal de l'internet français à la Bibliothèque nationale de France

Nous avons rencontré le 4 février 2013 Clément Oury[1], chef du service[2] du dépôt légal numérique, fondé en 2008 au sein du département du Dépôt légal de la Bibliothèque nationale de France. Il nous en a expliqué les principes et le fonctionnement.

Archivage du web, ou dépôt légal de l'internet français ?

L'expression courante d'« archivage du web », qui est la traduction de l'anglais *web archiving*, n'est pas pleinement appropriée pour qualifier la mission du service du dépôt légal numérique. Il convient plutôt de parler, avec le législateur, de « dépôt légal de l'internet français ». Dépôt légal, car cette mission s'attache, dans la continuité des activités bibliothéconomiques, à des publications – elle se rapproche cependant de l'archivistique en refusant de constituer une collection sur un critère de contenu, et en cherchant à recueillir de façon systématique et raisonnée l'activité d'un certain nombre de producteurs (en l'occurrence, les auteurs et éditeurs de sites internet). Surtout, le service consacré au numérique est le dernier né d'une série de structures assurant la mission du dépôt légal tel qu'il a été institué par l'édit du 28 décembre 1537 par François Ier et qui consiste à assurer la collecte, la conservation et la consultation de documents de toute nature, afin de constituer une collection de référence,

1. Nous remercions Clément Oury pour l'ensemble des éléments d'information et de réflexion qu'il nous a fournis.

2. Neuf personnes travaillent dans ce service.

Évelyne Cohen et Julie Verlaine, « Le dépôt légal de l'internet français… »,
S. & R., n° 35, printemps 2013, p. 211-218.

élément essentiel de la mémoire collective du pays. Après les imprimés, ce sont successivement les estampes, cartes et plans (1672), les partitions musicales (1745), les photographies et phonogrammes (1925), les affiches (1941), les vidéogrammes et documents multimédias composites (1975) puis les multimédias, logiciels et bases de données (1992), et enfin Internet (2006) qui sont soumis à l'obligation de dépôt légal[3].

Par la loi DAVDSI du 1er août 2006, entrée en application par le décret du 19 décembre 2011, est mis en place un dépôt légal de l'internet français, couvrant l'ensemble des « signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique[4] ». Cela concerne donc, en théorie, toute information circulant sur l'internet : aussi bien les *newsletters* diffusées par voie électronique, que tout site internet dont le nom de domaine se termine par « .fr », ou dont l'extension se rapporte au territoire français (« .re » pour la Réunion, « .nc » pour la Nouvelle-Calédonie…), ou dont l'auteur est une personne physique ou morale hébergée en France, ou bien enfin dont le contenu est produit sur le territoire national. Comme le souligne Valérie Game[5], le domaine ainsi couvert est celui ressortant de la justice française. La rupture avec la logique de support, qui a guidé les précédentes extensions du dépôt légal, est nette : avec l'internet, support dématérialisé s'il en est, une logique d'émetteur s'est imposée, en raison notamment de l'interpénétration des médias sur la plupart des sites. En pratique, le dépôt légal est surtout appliqué au domaine français (extension en « .fr »). Un accord entre la BnF et l'Association française pour le nommage Internet en coopération (AFNIC), qui fournit une liste exhaustive des sites en « .fr ».

Le décret de 2011 charge de la mission du dépôt légal de l'internet français l'Institut national de l'audiovisuel (Ina)[6] et la Bibliothèque nationale de France (BnF). Le choix de ces deux institutions est justifié par des impératifs de continuité de leurs missions documentaires : qu'il s'agisse de l'audiovisuel ou de l'imprimé, le mouvement de dématérialisation des collections est lié à la

3. « Qu'est-ce que le dépôt légal ? », page du site de la Bibliothèque nationale de France (http://www.bnf.fr/fr/professionnels/depot_legal_definition/s.depot_legal_mission.html, consulté le 4 février 2013).

4. Loi du 1er août 2006 et loi nº 2006-961 relative au droit d'auteur et aux droits voisins dans la société de l'information.

5. Valérie Game, Gildas Illien, « Le dépôt légal d'Internet à la Bibliothèque nationale de France : cadre juridique, modèle de collecte, évolutions des métiers », *Bulletin des bibliothèques de France*, nº 3 (2006), p. 82-85 (http://bbf.enssib.fr/consulter/bbf-2006-03-0082-013, consulté le 4 février 2013).

6. Sur l'Ina et le web lire Claude Mussou, « Et le Web devint archive : enjeux et défis », *Le Temps des médias*, nº 19 (2/2012), p. 259-266. Voir aussi http://www.institut-national audiovisuel.fr/actualites/webzine/depotlegalweb.html.

multiplication des contenus numériques, diffusés sous forme multimédias et sur le web mondial. L'Ina a donc la charge des sites possédés, animés, ou émis par une chaîne de radio ou de télévision, soit environ 9 000 sites[7] ; la BnF se charge du reste, soit près de 2,5 millions de sites.

Une collecte large et des collectes ciblées

Ce sont des « robots d'archivage » qui assurent la collecte des contenus sur le web. Ces logiciels, qui fonctionnent à la manière des logiciels de moteurs de recherche, disposent d'une liste d'adresses de pages internet (ou URL, *Uniform Resource Locator*) auxquelles ils se connectent et qu'ils archivent, et dont ils extraient et suivent les liens pour ensuite archiver les pages ainsi reliées. Il est possible de paramétrer les instructions de façon très fine. Plus le nombre de sites collectés est limité, plus le robot peut archiver des liens en profondeur.

Plutôt que de « site internet », les spécialistes[8] préfèrent parler d'*« entité* web[9] » tant les contours sont flous et les définitions poreuses. Leur objectif n'est pas l'exhaustivité, mais la représentativité[10] : c'est ce qui les pousse à privilégier l'idée de « collecte » (traduction officielle du terme anglais *crawl*), afin de mettre en avant le mouvement spécifique à 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.

La collecte large ou à grande échelle, qui remplit *stricto sensu* l'obligation de dépôt légal, est annuelle : les contenus de quelque 2,5 millions de sites internet du domaine national sont collectés automatiquement par les robots. Les archives rassemblées selon cette méthode représentent une « photographie instantanée » d'un ensemble de sites qui sont explorés superficiellement (à quatre ou cinq « clics » de profondeur depuis la page d'accueil). Chacun de

7. Le périmètre de l'Ina, parmi les sites français : sites émanant des services des médias audiovisuels ; web TVs et web radios ; sites principalement consacrés aux programmes radio et télé ; sites des organismes de l'environnement professionnel et institutionnel du secteur de la communication audiovisuelle.

8. Comme ceux du Medialab de Sciences po. http://www.medialab.sciences-po.fr/.

9. Paul Giard, « HyperText Corpus Initiative : how to help researchers sieving the web ? », proposition pour le panel « Using Web Archives panel », conférence lors de l'assemblée générale du consortium IIPC, 9 mai 2011 (http://www.medialab.sciences-po.fr/publications/Girard-HCI.pdf, consulté le 11 février 2013).

10. Dans le cas de la BnF spécialement. L'Ina ou le Médialab qui travaillent sur des corpus beaucoup plus réduits, peuvent tendre à l'exhaustivité.

ces instantanés annuels rassemble plusieurs centaines de millions de fichiers, représentant un poids croissant : de 2,5 téraoctets en 2004 à 33 téraoctets en 2012. Ils offrent une bonne représentation de la diversité et de la richesse des contenus de l'internet français, rendant compte notamment de la multitude de « petits » sites (une ou deux pages), principalement marchands ou publicitaires, qui sont actifs sur la toile. Mais ils ne peuvent rendre compte de tous les changements et mises à jour qui interviennent à l'intérieur de chaque site, dans un environnement éditorial souvent éphémère.

À cette première modalité s'ajoutent des stratégies de collectes moins automatiques, plus ciblées, sur un ensemble de 30 000 sites environ choisis par des spécialistes du web, bibliothécaires de la BnF et d'ailleurs, et chercheurs. Certaines de ces cibles documentaires sont collectées chaque jour, parfois plusieurs fois par jour pour celles dont les contenus sont les plus évolutifs (les sites d'information, tout particulièrement) ; d'autres, riches en ressources, sont collectées très en profondeur (sites institutionnels, comme celui du CNRS ou de l'Union européenne) ; d'autres enfin font l'objet d'une campagne de collecte sur une période temporelle définie (sites de festivals, de salons et autres manifestations éphémères ; événements politiques, comme campagnes électorales ; événements sportifs). Pour rendre plus efficaces ces campagnes ciblées, le service du dépôt légal de la BnF s'appuie sur un réseau interne de près de 80 correspondants répartis dans les différents départements (Littérature et art, Cartes et plans, Estampes et photographies…).

La BnF développe également de plus en plus un réseau externe, constitué des grandes bibliothèques chargées du dépôt légal imprimeurs, qui sont réparties dans chaque région de France. Elles sont associées au dépôt des contenus, lors de collectes portant sur des événements à fort impact régional. En 2012, pour les élections législatives, un réseau de vingt bibliothèques du dépôt légal imprimeurs (BDLI), dont trois outre-mer, a opéré.

La Bibliothèque nationale de France a lancé plusieurs opérations successives d'archivage des sites électoraux. Les premières campagnes en ligne qu'elle a couvertes furent celles des élections présidentielle et législatives de 2002.

Ce principe de constitution de sources primaires n'est pas, explique Clément Oury[11], pour [la BnF] une nouveauté : ses agents recueillent depuis le XIXe siècle le matériel de propagande électorale (tracts, affiches…). Mais il a fallu s'adapter au caractère dématérialisé de cette documentation : au fil des campagnes, la BnF a su progressivement élaborer une politique stable de sélection et d'archivage, constituer des groupes d'experts, associer un nombre

11. Clément Oury, « Soixante millions de fichiers pour un scrutin », *Revue de la BnF,* 40, 2012/1, p. 84-90.

croissant de partenaires à son activité. Il lui a également fallu mettre en place et améliorer les outils de collecte et d'accès aux collections. Pour faire face à l'enjeu que constituait la capture du web électoral, espace vaste et en perpétuelle expansion, elle a dû prendre en compte les risques, mais aussi s'appuyer sur les possibilités que lui offrait ce média.

Les sites liés à l'élection présidentielle et aux élections législatives de 2012, par exemple, ont été capturés sur une période de huit mois, de janvier à juillet 2012. Le projet a mobilisé plus d'une soixantaine d'agents, à la BnF ainsi que dans les vingt bibliothèques en région – de métropole comme d'outremer. Il a permis la capture de plus de 10 500 sites ou parties de sites, à des fréquences régulières (allant de quatre fois par jour à une seule fois, selon les sites). Tous les types d'acteurs du débat politique sur la Toile ont été représentés : sites de candidats, de partis ou d'organisations de soutien, mais aussi presse en ligne, blogs de militants, réseaux sociaux, observatoires de la « Net-politique ». La collection constituée représente un ensemble de 380 millions de fichiers, soit 11 téraoctets de données.

Pour quels usages ?

Une fois collectés et archivés sur les serveurs de la BnF, les publications numériques sont mises à disposition des chercheurs accrédités pour la bibliothèque de recherche sur des postes informatiques installés en salle de lecture. C'est le grand paradoxe du dépôt légal du web en France : les données qu'il collecte sont en ligne, donc universellement et librement accessibles, mais la consultation après archivage est soumise à condition et ne peut se faire qu'entre les murs de la bibliothèque. Il y a à cette politique plusieurs raisons, à commencer par le respect du droit de la propriété intellectuelle et par le souci de suivre les recommandations de la CNIL[12] en matière de protection des données personnelles, en rendant inaccessibles les contenus sans accréditation.

Si le nombre de consultations des archives des sites internet est aujourd'hui assez faible (une cinquantaine par mois), c'est parce que la plupart des chercheurs travaillant sur le web n'ont pas acquis le réflexe d'en consulter les archives en bibliothèques. Mais le développement de l'histoire des nouveaux médias, conjugué à la marginalité des pratiques d'auto-archivages des sites rendront précieux, à terme, les documents collectés par la BnF qui constituent la mémoire de la toile française. Les contenus les plus anciens remontent à

12. Commission nationale de l'informatique et des libertés.

Évelyne Cohen et Julie Verlaine, « Le dépôt légal de l'internet français… »,
S. & R., n° 35, printemps 2013, p. 211-218.

1996 : ils ont été donnés par la fondation américaine Internet Archive, une immense bibliothèque numérique spécialisée dans l'archivage du web, fondée à San Francisco en 1996[13]. Depuis 2002, les bibliothécaires de la BnF ont mené une politique d'archivage du web, d'abord à titre expérimental puis, à partir de 2006, de manière systématique et intégrée à la politique documentaire de l'institution.

Sur les postes de consultation installés en salle de lecture, trois modes d'interrogation du corpus sont possibles à partir de l'interface d'accès. Une première grille d'interrogation se fait par adresse ou URL, puis par date. Une recherche portant sur le site d'information et de débat « Rue89.com », par exemple, renvoie à plus de 2 500 versions archivées, entre 2007 et 2012. La navigation peut se faire dans l'espace (par le suivi des liens), et dans le temps (de collecte en collecte). La seconde grille, encore à l'état expérimental, propose une recherche par mots-clefs, et la troisième des « parcours guidés », autant de voyages dans les collections permettant d'en présenter la richesse.

Plusieurs thématiques de recherche ont été ou sont explorées à travers des collectes ciblées. Ainsi tout « chercheur[14] » qui « exprime un besoin légitime » d'explorer un thème de recherche peut s'adresser au service et demander à ce que soient collectés des sites en rapport avec sa recherche. Les départements de la BnF ont engagé des collaborations avec des laboratoires de sciences humaines et sociales qui travaillent, par exemple, sur les impacts politiques et sociaux des sports[15], ou encore avec l'Association pour l'autobiographie et le patrimoine autobiographique dirigée par Philippe Lejeune[16].

À terme, l'objectif recherché est de pouvoir utiliser et partager les métadonnées des publications numériques. À travers le consortium international pour la préservation de l'internet IIPC[17], dont la BnF est membre fondateur, se développent des pratiques de coopération entre une quarantaine d'institutions dans le monde qui échangent sur les normes, les outils, les pratiques[18]. Il faut remarquer que ces pratiques de coopération dessinent un espace d'échanges principalement européens, nord américains et dans une moindre mesure

13. Internet Archive (http://archive.org/, consulté le 4 février 2013). Dès 1996, un ancien du MIT, Brewster Kahle, a eu la conviction qu'il fallait protéger l'internet de l'oubli.

14. Au sens large et en fonction de besoins universitaires, professionnels ou personnels.

15. Comme le laboratoire PACTE à Grenoble (http://www.pacte-grenoble.fr/).

16. Voir « Du journal intime au blog », http://blog.bnf.fr/lecteurs/index.php/2009/04/06/du-journal-intime-au-blog/.

17. Gildas Illien, « Une histoire politique de l'archivage du web », *BBF*, 2011/2, p. 60-68 (http://bbf.enssib.fr/, consulté le 5 février 2013).

18. Voir http://www.netpreserve.org/.

asiatiques ; et qu'en sont absentes l'Afrique (à l'exception de la bibliothèque alexandrine) et l'Amérique du Sud.

Les missions fondamentales du consortium IIPC sont de :

- travailler en collaboration, dans le cadre législatif de leurs pays respectifs, pour identifier, développer et faciliter la mise en œuvre de solutions permettant de sélectionner, de collecter et de préserver les contenus de l'internet et d'en assurer l'accessibilité ;
- faciliter la couverture internationale des collections d'archives de contenus de l'internet, en conformité avec leurs cadres législatifs nationaux et en accord avec leurs politiques respectives de développement des collections nationales ;
- plaider vigoureusement au niveau international en faveur d'initiatives et de lois encourageant la collecte, la préservation et l'accès aux contenus de l'internet.

La constitution de la fédération des pratiques de chercheurs sur l'archivage du web s'inscrit dans cette perspective.

La coopération internationale entre archivistes du web s'effectue sur plusieurs plans, même si chaque pays a ses propres réglementations et donc sa propre pratique de la collecte. La British Library, par exemple, ne possèdait pas le dépôt légal et devait donc demander aux ayants droit des sites l'autorisation de collecter leur site jusqu'à l'entrée en vigueur de la loi légale Deposit Act, le 6 avril 2013. Ce qui permet en retour de mettre en ligne les contenus.

Certains sujets font l'objet de collectes internationales. Cela a été le cas pour les Jeux olympiques de Londres en 2012. Grâce au système Memento[19], on peut aujourd'hui échanger les métadonnées de ces collections d'archives du web et à terme il sera possible d'avoir une interface commune.

L'avenir

Des enquêtes[20] ont été menées auprès des chercheurs afin de déterminer au mieux leurs attentes[21] quant à l'archivage du web. Il semble important de tracer la frontière entre archives personnelles et archives publiées donc publiques. Ainsi la page Facebook ou le site Twitter d'un homme politique est-il ouvert :

19. Voir http://mementoweb.org/.

20. Voir http://www.bnf.fr/documents/enquete_archives_web.pdf.

21. BnF, archives de l'internet : représentations et attentes des utilisateurs potentiels. Étude réalisée fin 2010-début 2011.

on peut le considérer comme une publication qui a vocation à être archivée. D'une certaine manière, même si certains usagers n'en ont pas clairement conscience, tout écrit sur le web est public mais il reste soumis au droit d'auteur. Il semble donc particulièrement utile d'opérer la distinction entre mémoire publique, mémoire commerciale et mémoire personnelle. En instituant le dépôt légal du web l'État a 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 » (Clément Oury)*.

* Pour continuer, consultez la bibliographie en ligne : http://www.bnf.fr/documents/bibliographie_dl_web.pdf.

Évelyne Cohen et Julie Verlaine, « Le dépôt légal de l'internet français… »,
S. & R., n° 35, printemps 2013, p. 211-218.