Le dépôt légal de l'internet en pratique :

→ LES MOISSONNEURS DU WEB

Gildas Illien
Bibliothèque nationale de France
gildas.illien@bnf.fr

Gildas Illien *pilote le projet du dépôt légal du web depuis 2005 et dirige aujourd'hui le nouveau service de la BnF chargé du dépôt légal numérique. Depuis 2007, il est également responsable technique et trésorier du Consortium international pour la préservation de l'internet. Il a exercé auparavant à la bibliothèque universitaire de Paris-VIII, a fait partie de l'équipe de préfiguration de la bibliothèque de l'Institut national d'histoire de l'art et a dirigé les médiathèques de l'Institut français de Vienne et du centre culturel français d'Oslo. Il est l'auteur d'un essai,* La place des arts et la révolution tranquille : les fonctions politiques d'un centre culturel *(Presses de l'université Laval, 1999) et de plusieurs articles sur l'évaluation des bibliothèques, les services aux publics et l'archivage du web. Conservateur des bibliothèques depuis 2003, il est aussi diplômé de Sciences Po Paris et titulaire d'un master en communication de l'université McGill (Montréal).*

Début 2006, le *Bulletin des bibliothèques de France* invitait la Bibliothèque nationale de France (BnF) à faire part à ses lecteurs des actions engagées pour étendre sa mission de dépôt légal aux publications de l'internet [2]. Nous indiquions alors les principes directeurs, mais aussi les incertitudes qui pesaient sur ce qu'il convenait encore d'appeler un projet expérimental. Un peu plus de deux ans plus tard, de nombreux obstacles ont pu être levés : le législateur a donné à l'archivage du web un cadre juridique ; l'établissement s'est doté d'une infrastructure informatique assurant la sauvegarde et le traitement continus de grandes quantités de données ; les collections sont désormais accessibles au public ; enfin, l'équipe projet s'est muée en un service du dépôt légal numérique qui a trouvé sa place dans l'organisation de la BnF.

Comment ce dispositif fonctionne-t-il concrètement ? Préfigure-t-il des évolutions profondes pour la profession ? En particulier, quelle marge de manœuvre et quelle valeur ajoutée le bibliothécaire peut-il conserver à l'heure de la massification des collections nées numériques ? Après avoir précisé le mandat et le champ d'intervention de la BnF dans ce domaine, on présentera les processus de collecte et de valorisation qui caractérisent les changements – mais aussi les continuités – liés à l'apparition d'un web patrimonial.

Le champ des possibles : nouveautés juridiques, continuités patrimoniales

Résumons ce que le dépôt légal de l'internet recouvre exactement en insistant sur les points que la législation et la pratique des dernières années ont permis d'éclaircir[1].

Au terme d'une aventure juridique qui a duré près de sept ans, c'est finalement sur le support législatif offert par la loi relative au droit d'auteur et aux droits voisins dans la société de l'information[2] qu'a été introduit le dépôt légal de l'internet. Celui-ci a intégré les dispositions du Code du patrimoine (articles L 131-1 et suivants). La définition du champ d'application de cette nouvelle extension du dépôt légal n'allait pas de soi. Le dispositif devait en effet concilier des contraintes fortes, telles que le droit de la propriété intellectuelle ou la protection des données personnelles, et les caractéristiques du web, dont l'échelle, la temporalité, l'hyper-connectivité et le caractère international constituaient des défis jusqu'ici inconnus des institutions de mémoire.

La collecte des sites se faisant au moyen de procédures largement automatisées, le législateur a d'abord

1. Voir [2] et [3] pour une description détaillée des aspects techniques et juridiques du dépôt légal d'internet à la BnF et [10] pour un survol des stratégies d'archivage du web à travers le monde.
2. Titre IV de la loi dite Dadvsi n° 2006-961 du 1er août 2006, en ligne sur Legifrance : www.legifrance.gouv.fr/affichTexte.do?cidTexte = JORFTEXT000000266350 & dateTexte

adopté une approche pragmatique, capable de couvrir l'immensité du web sans entraver le fonctionnement des robots par la multiplication d'opérations manuelles. En particulier, la loi autorise les institutions mandataires à copier les sites web sans solliciter l'autorisation préalable des éditeurs. Cette exception au droit d'auteur a été déterminante dans la définition du modèle de production de la BnF : les opérations s'avéreraient économiquement très difficiles dans d'autres conditions, comme le démontrent plusieurs expériences étrangères[3].

La contrepartie de cette exception est la restriction des conditions de communication des archives : le projet de décret d'application devrait en effet limiter la consultation aux chercheurs accrédités, dans les emprises des établissements mandataires (afin de protéger le droit d'auteur), sur des postes individuels de consultation mis à disposition des usagers (afin d'empêcher des traitements de masse qui iraient à l'encontre de la protection des données personnelles).

L'internet français

Comment définir l'internet français ? Alors que le dépôt légal français est historiquement lié aux notions de support (depuis l'ordonnance de Montpellier de 1537, le dépôt légal est décliné et réparti selon une typologie historiquement stratifiée par l'apparition des médias) et de territoire (est traditionnellement assujettie au dépôt légal toute publication publiée ou diffusée en France), il a fallu faire preuve d'imagination pour circonscrire ce champ. La loi n'apporte pas de réponse explicite, mais le décret en préparation devrait confirmer la pratique engagée par la BnF.

Une première mesure a consisté à abandonner la notion de support physique pour prendre en compte un réseau de documents immatériels pouvant revêtir une grande diversité de formes : les « *signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique* » (article L. 131-2 du Code du patrimoine) ; cette définition extrêmement large et évolutive permet de viser un grand nombre de fichiers, pourvu qu'ils fassent bien l'objet d'une communication au public (ce qui exclut du champ les correspondances privées ou les espaces privés hébergés sur le web).

Il convenait ensuite d'adopter pour le web une conception souple du territoire national (afin d'autoriser les techniques de captation exploratoires à grande échelle), mais qui reste applicable aux éditeurs de sites, notamment dans le cas où la BnF est conduite à réclamer à un éditeur les « codes et restrictions d'accès » de certains documents protégés, comme le prévoit la loi. Le projet de décret définit ainsi comme d'origine française les sites « *enregistrés sous le nom de domaine .fr ou tout autre nom de domaine enregistré auprès des organismes français chargés de la gestion de ces noms, et/ou produits sur le territoire français ou enregistrés par une personne domiciliée en France* ». Si l'on comprend ainsi que le noyau dur du champ d'application consiste en la liste des sites en .fr gérée par l'Afnic (Association française pour le nommage internet en coopération), le robot de collecte est néanmoins autorisé à capturer des fichiers hébergés sous un autre domaine de haut niveau (.com ou .org par exemple), pourvu que son éditeur soit basé en France ou qu'on puisse démontrer que l'œuvre en question a été produite sur le territoire.

Dans l'esprit de ce dispositif, les conditions n'ont pas à être vérifiées préalablement à la collecte (ce qui en invaliderait l'économie) mais peuvent être examinées a posteriori en cas de litige ou de réclamation. Cette conception du champ constitue un compromis pragmatique. Elle laisse à la BnF comme aux éditeurs la possibilité de vérifier selon des critères objectifs si la loi s'applique ou non à une publication. Elle permet en tout cas de collecter au-delà du seul .fr, dont on sait qu'il ne concerne qu'une portion limitée des contenus susceptibles d'intéresser le patrimoine national[4].

> "La conception du champ permet de collecter au-delà du seul .fr, dont on sait qu'il ne concerne qu'une portion limitée des contenus susceptibles d'intéresser le patrimoine national"

3. L'expérience de la British Library [7] est éloquente. La législation britannique oblige nos homologues d'outre-Manche à obtenir l'accord des ayants droit d'un site avant sa capture. Depuis 2004, sur 6 500 sites identifiés par une équipe de quatre personnes, seuls 1 800 ont pu être archivés (les autres n'ayant pas reçu l'accord des éditeurs), pour constituer une collection d'un volume total limité à 840 gigaoctets. La British Library estime que, sans modification de la loi, le dispositif actuel ne lui permettra pas de couvrir plus de 0,6 % du domaine britannique de l'internet d'ici dix ans.

4. Le. fr a connu un fort développement avec l'assouplissement progressif de ses règles d'attribution et notamment son ouverture aux particuliers en juin 2006. L'Afnic comptabilisait 1 205 734 noms de domaine en. fr à la date du 6 septembre 2008. Pour autant, une étude de l'Afnic [1] note que moins de 30 % des sites web « français » seraient hébergés en .fr. Des analyses conduites par la BnF sur des corpus thématiques confirment cette analyse (seulement 36 % des sites capturés lors des élections présidentielle et législatives de 2007 étaient hébergés en .fr).

Les limites du web profond

Peut-on cependant tout archiver ? Non, bien sûr. L'archivage du web s'appuie sur des techniques relativement primitives au vu des évolutions actuelles de l'édition en ligne. Conçues au tout début du millénaire, celles-ci comportent leurs propres limites techniques et économiques. Les fichiers en ligne sont identifiés et copiés automatiquement au moyen de robots moissonneurs comparables à des internautes automatiques : ils cliquent de lien en lien sur tous les sites qu'ils rencontrent.

À partir d'une liste d'adresses de sites, ils explorent le web de page en page et capturent les fichiers qu'ils découvrent, soit en profondeur (liens entrant à l'intérieur d'un même site), soit en largeur (liens sortant vers d'autres sites). C'est cette liste de départ, ainsi que les paramètres de moissonnage (fixant la profondeur, la date ou la fréquence des collectes), qui déterminent le contenu, la qualité et l'étendue de la collection. Les robots rencontrent régulièrement des obstacles et se heurtent aux limites du web profond.

De plus, une campagne de moissonnage est potentiellement illimitée : pour ménager les ressources de calcul et de stockage, on doit pouvoir la stopper sans qu'on ait l'assurance que tous les contenus aient été effectivement collectés. Pour ces raisons, les archives du web sont des documents lacunaires puisqu'il peut manquer des fichiers, des pages, mais aussi parce qu'il est impossible de moissonner tous les sites en permanence : les collections constituées sont rarement des séries exhaustives ; elles se présentent plutôt comme des recueils de traces ou d'échantillons du web liés entre eux comme dans leur environnement de consultation initial.

Un circuit documentaire pensé pour la masse

Le patrimoine né numérique constitué par la BnF depuis 2004 représente déjà 130 téraoctets de données, soit 130 millions de millions d'octets et 12 milliards de fichiers : c'est l'une des plus grandes collections d'archives du web au monde, après celles d'Internet Archive et de la Bibliothèque d'Alexandrie [5].

Face à de tels chiffres, la seule issue raisonnable est de rechercher l'automatisation maximale des traitements à toutes les étapes du cycle de vie du document. Le caractère éphémère du web implique par ailleurs une réactivité très forte [6]. Les problématiques du volume et de l'éphémère convergent ainsi vers le déplacement des traitements scientifiques et humains, les plus longs et les plus chers, de l'amont vers l'aval de la gestion de la collection, afin de limiter les coûts ainsi que le risque d'une disparition pure et simple des contenus.

Aucune gestion humaine ne peut donc être envisagée de manière systématique. Pour autant, on ne peut s'en remettre entièrement à des robots pour constituer le patrimoine de demain. On ne peut non plus déposséder les bibliothécaires de la gestion de collections qui constituent le prolongement de leurs missions.

Enfin, favoriser dès le début l'appropriation de ce nouveau type de collection par les professionnels est apparu à la BnF comme une condition essentielle de leur valorisation auprès du public. Le repérage, la sélection et la valorisation des sites à l'unité par des professionnels ont donc été organisés, mais réservés à des domaines prioritaires dans le cadre de projets circonscrits.

La collecte des sites est ainsi assurée par moissonnage automatique ou semi-automatique en associant ces deux approches. Nous détaillerons plus loin les processus métier liés à cette activité. La suite du circuit du document, dont nous ne donnons ici qu'un aperçu, procède d'une vision analogue, où l'intervention humaine est réservée aux procédures auxquelles elle ajoute une valeur.

Le contrôle qualité des archives se fait à l'issue ou directement en cours de collecte par échantillonnage et à l'aide de rapports statistiques portant sur de grands volumes. Ces rapports permettent d'analyser le poids, la taille ou la répartition des collections par type de fichiers, afin d'évaluer leur qualité et d'en dégager les principales caractéristiques [8]. On utilise également des outils de dédoublonnage ou de contrôle automatique qui vérifient par exemple la syntaxe d'une URL ou sa présence en ligne.

Les Petabox, baies de stockage sur disque dont la capacité varie entre 60 et 120 téraoctets. © BnF

5. L'Alexandrina bénéficie d'une coopération exceptionnelle d'Internet Archive, qui lui offre une copie de ses collections.
6. D'après un rapport du projet Planets, la durée de vie moyenne d'un site web serait de 44 jours. La BnF a calculé qu'un tiers des sites qu'elle a collectés lors de l'élection présidentielle de 2002 a totalement disparu.

Après collecte, les fichiers sont stockés dans des fichiers plus gros appelés ARC, un format container propre aux archives du web et qui facilite la manipulation de gros volumes de données en concaténant des enregistrements où fichiers compressés et métadonnées (date de capture, poids…) sont rassemblés[7]. Les fichiers ARC sont ensuite entreposés dans de vastes baies de stockage sur disque (les Petabox), dont la capacité varie entre 60 et 120 téraoctets. Ces baies sont destinées à l'exploitation et au stockage courants des collections, aux fins de consultation en particulier : c'est l'équivalent du libre accès.

Une infrastructure distincte est dédiée à la conservation des données – c'est le magasin. Les fichiers ARC font en effet l'objet d'une copie de sauvegarde dans le Système de préservation et d'archivage réparti (Spar) de la BnF[8]. Celui-ci devrait devenir d'ici 2010 l'entrepôt intelligent chargé d'assurer leur archivage pérenne en conformité avec le modèle conceptuel de l'OAIS (Open Archival Information System). Les processus préalables d'identification et de validation des formats de fichiers sont des étapes qui s'annoncent complexes, compte tenu de la quantité et de la diversité des données. Pour qu'elles restent lisibles sur le long terme, les données feront ensuite l'objet d'opérations de migration ou d'émulation. Une coopération internationale a été engagée en 2008 afin d'étudier les meilleurs scénarios.

Le signalement des données répond lui aussi à une logique de masse. Le catalogage des sites archivés a été exclu au profit de l'indexation automatique. Toutes les collections de la BnF sont aujourd'hui indexées pour une recherche par URL avec le logiciel Wayback Machine : on peut retrouver un site si l'on connaît son adresse exacte. Le délai moyen entre la capture d'un site et son indexation est actuellement de l'ordre de deux semaines. Et c'est cette indexation qui permet de naviguer dans les archives comme sur le web vivant, en cliquant de lien en lien. Afin de faciliter les allers-retours entre les archives et l'internet (une démarche utile pour retrouver les adresses des sites en interrogeant un moteur de recherche en ligne), la BnF a mis en place une interface qui autorise la circulation entre les deux espaces (web mort et web vivant), tout en assurant leur cloisonnement, afin d'éviter la confusion entre les temporalités de publication.

La boîte à outils de l'archivage du web

Tous les logiciels utilisés par la BnF sont des logiciels libres développés collaborativement par Internet Archive et les bibliothèques membres du consortium IIPC.

Heritrix, logiciel de moissonnage : http://crawler.archive.org

Wayback Machine, logiciel d'indexation et interface de navigation pour la recherche par URL
http://archive-access.sourceforge.net/projects/wayback

NutchWAX, logiciel expérimental d'indexation plein texte pour la recherche par mot
http://archive-access.sourceforge.net/projects/nutch

(W) ARC : le format container des archives du web

ARC : www.archive.org/web/researcher/ArcFileFormat. php

WARC : http://bibnum.bnf.fr/WARC

L'indexation plein texte reste un objectif prioritaire pour permettre la recherche par mot. Elle s'avère en effet l'outil le plus intuitif et le plus attendu des utilisateurs, qui veulent naturellement rechercher dans les archives comme ils le font sur le web. Le logiciel NutchWAX est encore expérimental, et moins de 5 % des collections de la BnF ont pu être indexés en plein texte. Compte tenu des performances et des coûts d'indexation actuels, on envisage de réserver, dans un premier temps, ce mode d'indexation aux collections issues d'une sélection humaine, afin de les valoriser.

D'autres opérations de signalement sont en cours d'expérimentation. Pour celles-ci, on le verra, l'intervention du bibliothécaire peut être sollicitée. Cette contribution est cependant davantage éditoriale (valorisation de corpus, thèmes, actualités ou parcours) que descriptive. Elle s'inscrit dans l'évolution plus globale de la bibliothèque éditrice, c'est-à-dire agissant en tant qu'architecte d'interfaces et de services plutôt que comme opérateur direct du traitement des données.

À terme, au-delà de la consultation des données, la BnF souhaite ainsi s'impliquer dans la recherche-développement d'outils d'exploitation et d'analyse de gros volumes (fouille de données, cartographies dynamiques du web…). La structure de la collection constituée permet en effet de retrouver non seulement des sites web, mais aussi des informations précises sur leur contexte de publication, les relations entre sites, ainsi que l'évolution de ces relations dans le temps et la géographie du web (ce qu'on appelle le *link mining*). C'est un secteur particulièrement prometteur, notamment pour la nouvelle sociologie du web, où des développements importants sont attendus.

Plus près de nous, le cercle vertueux du circuit documentaire a enfin atteint son objectif : le public. Ceux que nous proposons d'appeler les archinautes ont commencé à consulter les collections du dépôt légal de l'internet en avril 2008. Ils sont encore peu nombreux (quelques dizaines), mais les premières questions et les premiers sujets de recherche ont surgi l'été dernier en salle de lecture. Une docto-

Logo BnF des archives de l'internet

7. [3] Le format ARC devrait bientôt laisser place au format WARC, en cours de normalisation à l'ISO.

8. Voir à ce sujet, dans le présent numéro du *BBF*, l'article d'Emmanuelle Bermès, Marie-Élise Fréon et Frédéric Martin, « Tous les chemins mènent au numérique : éclairage sur l'activité de la Bibliothèque nationale de France », p. 34-39.

rante en linguistique qui travaille sur la parole des femmes est venue spécialement d'Italie pour étudier le blog, disparu, de Marie-Georges Buffet. Des chercheurs en sciences politiques ont consulté les archives pour étayer leur recherche d'information sur des sujets comme « l'utilisation d'internet pendant la campagne présidentielle de 2007 » ou encore « la logique de recherche d'information des citoyens autour de la question européenne ». D'autres questions, parfois insolites, nous ont été posées : « *Est-il possible de déposer son site web pour qu'il soit archivé ?* » ou encore « *Puis-je retrouver dans les archives le site* [disparu] *de Jallier, maison de faïence à Moustiers ?* ». C'est ce premier contact, tant attendu, qui va permettre de faire évoluer la politique d'enrichissement des fonds comme l'ergonomie et les fonctionnalités des outils.

Le déploiement de l'interface de consultation des archives (actuellement installée sur une centaine de postes des sites François-Mitterrand et Richelieu) dans l'ensemble des salles de recherche de la BnF est prévu pour le printemps 2009. Si le décret en préparation confirme cette possibilité, nous espérons étendre ensuite cet accès aux bibliothèques de dépôt légal imprimeur en région. Une politique de communication, qui cible prioritairement les communautés scientifiques susceptibles d'être intéressées par ce nouveau matériau, a été engagée. Un dispositif d'évaluation quantitative et qualitative de la consultation (statistiques, entretiens, enquête en ligne, journaux tenus par les bibliothécaires en service public) a aussi été mis en place afin d'en tirer un premier bilan à l'été 2009.

Un dispositif industriel pour les collectes massives du domaine national

À la BnF, les collectes larges du domaine national permettent d'engranger en quelques semaines une dizaine de téraoctets de données. Depuis 2004, quatre collectes larges ont ainsi été réalisées par Internet Archive depuis la Californie, indexées puis livrées à la BnF sur machine (les Petabox). La prochaine est programmée en novembre 2008. La Bibliothèque a complété ces envois par des acquisitions rétrospectives extraites des collections mondiales d'Internet Archive – ces incunables du web remontent à 1996. Collectes du domaine français et acquisitions rétrospectives pèsent aujourd'hui respectivement 21 et 70 téraoctets de données, soit près de 80 % des collections web de la BnF. On comprend que, sans le recours à Internet Archive, la Bibliothèque n'aurait pu sauvegarder seule une aussi grande quantité de données en si peu de temps. L'internalisation complète par la BnF de ses collectes larges du domaine national constitue toutefois une priorité et son principal défi à l'horizon 2009.

Depuis l'an passé, ces collectes exploratoires sont lancées à partir de la liste des noms de domaines que l'Afnic communique à la BnF dans le cadre d'une convention signée en 2007[9]. Les collections constituées dans ce cadre répondent à la tradition française du dépôt légal, qui doit rester aveugle, au sens où il se veut plus représentatif que qualitatif. Elles contiennent ainsi nombre de contenus qu'un bibliothécaire n'aurait pas forcément repérés ni choisis (blogs et autoédition, sites commerciaux, publicitaires, pornographiques...). Cela garantit le caractère à la fois massif et représentatif du patrimoine constitué dans la continuité de notre dépôt légal, qui se veut un « miroir » de tout ce que la communauté nationale publie et non une sélection de ses meilleures œuvres, l'appréciation étant laissée aux générations futures.

Le pilotage des opérations de collecte et son extension prochaine à des volumes plus conséquents ont impliqué le recrutement de sept agents. Il a fallu plusieurs années pour constituer cette équipe qui s'est en grande partie formée sur le tas, en tirant parti des opportunités de collaboration qu'offre le Consortium international pour la préservation de l'internet (IIPC). En effet, la supervision des robots de collecte n'est pas une opération purement technique. La qualité d'une collection d'archives se détermine pendant la capture des sites, dont la collecte se programme, se surveille, s'évalue. Les ingénieurs et techniciens impliqués dans ce travail doivent être pleinement conscients des contenus qu'ils produisent, des enjeux documentaires et des risques patrimoniaux qui y sont associés. Les bibliothécaires qui sont leurs interlocuteurs ne peuvent quant à eux ignorer le fonctionnement des serveurs ni les coûts informatiques qui pèsent sur la production au risque de dérégler toute l'économie et la cohérence du travail.

> “Ceux que nous proposons d'appeler les archinautes ont commencé à consulter les collections du dépôt légal de l'internet en avril 2008”

C'est pourquoi l'équipe technique est répartie entre le département des Systèmes d'information et le département du Dépôt légal de la BnF. Les informaticiens, chargés de la maîtrise d'œuvre, dialoguent en permanence avec les bibliothécaires du dépôt légal. Ceux-ci assurent la maîtrise d'ouvrage des collectes ainsi que l'interface avec tous les bibliothécaires de la BnF commanditaires de contenus web. Ils centralisent leurs commandes et évaluent leur impact « économique »

9. On trouvera en [9] un bilan approfondi des expériences conduites par la BnF avec Internet Archive pour définir son cœur de cible « national » dans le cadre des collectes larges depuis 2004 ainsi qu'un compte rendu de l'utilisation de la liste des noms de domaines fournie par l'Afnic à partir de 2007.

(coût de production humain et machines) avant de planifier et de gérer au quotidien la production. À mesure que son carnet de commandes se remplit, l'équipe prend conscience qu'elle développe des compétences hybrides, où le documentaire et l'informatique s'entremêlent de plus en plus. Pour ces bibliothécaires d'un genre nouveau, on a inventé la fonction de « chargé de collections numériques ».

Les nouveaux conservateurs du web au service des collectes ciblées

Pour les collectes ciblées qui viennent compléter la manne des instantanés annuels, la BnF a mis en place une organisation de coopération interne qui s'élargit progressivement à des partenaires extérieurs. Ce dispositif complémentaire des collectes du domaine national vise à apporter une valeur ajoutée à la collection constituée par robot ; il s'avère également essentiel pour assurer la valorisation des fonds auprès du public des chercheurs. Chacun des grands départements de l'établissement responsable d'un ensemble de disciplines ou de supports s'est doté d'une équipe de correspondants du dépôt légal de l'internet et définit, chacun dans son domaine, les axes de prospection prioritaire à partir d'une méthodologie et d'une terminologie communes.

Des priorités ont pu ainsi être établies, qu'il s'agisse de secteurs bien connus pour lesquels le web pose clairement un enjeu de continuité patrimoniale (par exemple : les encyclopédies, les publications officielles, les périodiques, la musique ou les films en ligne), ou de domaines nouveaux tels que les sites du web 2.0 (sites de réseaux sociaux, blogs, wikis...) et du Net-Art, dont l'archivage est indispensable à la compréhension du web et de son appropriation par la société et les créateurs français. C'est dans le cadre de ces travaux que la BnF a par exemple lancé en 2007 une opération de collectes ciblée sur les journaux intimes en ligne (projet conduit en partenariat avec l'Association pour la pensée autobiographique[10]), ou encore les vidéos du site de partage Dailymotion.

À côté de ce travail de fond, qui vise à assurer l'élargissement progressif et pérenne de tous les services d'entrées de la BnF (dépôt légal, acquisitions, coopération...) aux contenus du web, l'établissement soutient des projets spécifiques qui favorisent la mobilisation des équipes et des partenaires autour d'opérations fortes mais limitées dans le temps. La constitution de ces corpus vise à réaliser des produits d'appel et des clés de valorisation au sein d'une collection si importante en volume qu'elle nécessitera des points d'entrée intelligibles et attractifs à destination des premiers archinautes. Après les collectes électorales expérimentales de 2002 et 2004, le projet pilote « Internet en campagne », centré sur la collecte des sites électoraux des élections présidentielle et législatives de 2007 a ainsi joué un rôle de locomotive pour l'internalisation de toute la chaîne de production.

Le consortium IIPC

Histoire et missions

IIPC (International Internet Preservation Consortium) a été fondé en 2003 à l'initiative de 10 bibliothèques nationales d'Amérique du Nord, d'Australie et d'Europe (dont la BnF) et de la fondation américaine Internet Archive. Le consortium a pour principaux objectifs la promotion de l'archivage du web dans le monde et le développement collaboratif de logiciels libres et de normes pour la collecte, la consultation et la préservation à long terme du patrimoine de l'internet.

Composition

En 2007, IIPC a entrepris son élargissement à de nouvelles institutions. Il compte aujourd'hui 37 membres parmi lesquels :

- en Amérique du Nord, la Bibliothèque du Congrès, Bibliothèque et archives Canada, Bibliothèque et archives nationales du Québec et la California Digital Library ;
- en Europe occidentale, l'Institut national de l'audiovisuel, la British Library, les Archives nationales britanniques, toutes les BN des pays nordiques, celles d'Allemagne, de France, d'Italie, de Suisse, de Catalogne, des Pays-Bas, la fondation European Archive et la société Hanzo Ltd ;
- en Europe orientale, les BN d'Autriche, de Croatie, de Slovénie, de la République tchèque, de Lettonie, de Pologne ;
- au Moyen-Orient, la BN d'Israël ;
- en Extrême-Orient, les BN du Japon, de Chine, de Singapour et de Corée du Sud ;
- et en Australasie, la BN de Nouvelle-Zélande.

Gouvernance

La présidence a été assurée par la France jusqu'en 2007 puis par l'Islande en 2008.
Le comité de pilotage est constitué des membres fondateurs d'IIPC.
Le comité technique regroupe des experts internationaux chargés de l'agenda technique d'IIPC.
La BnF assure le pilotage technique et administratif ainsi que la trésorerie.
La Bibliothèque du Congrès est chargée de la communication et des services aux membres.
La première assemblée générale d'IIPC s'est tenue à Paris en 2007, la deuxième à Canberra en 2008, et la troisième se réunira à Ottawa en 2009.

En savoir plus : www.netpreserve.org

Internet en campagne

La collection issue de cette mobilisation sans précédent contient 5 800 sites sélectionnés par une équipe de 24 agents sollicités pendant huit mois, en partenariat avec 8 bibliothèques de dépôt légal imprimeur en région[11]. Ce fonds unique représente un volume de 63 millions de fichiers, soit 3,4 téraoctets de données. Il présente aussi bien les sites officiels des candidats et formations politiques qu'un échantillon des blogs de la société civile et des « journalistes citoyens » en cours de campagne. À l'ouverture de la communication des archives au public en avril 2008,

10. Voir le site de l'APA http://sitapa.free.fr et l'article de Bernard Massip signalé en référence [11].

11. Suite à un appel à partenariats lancé en 2006 par la BnF, les BDLI de Caen, Dijon, Lille, Lyon, Limoges, Nouméa, Poitiers et Strasbourg ont rejoint ce projet pour assurer la couverture de la « web campagne » des élections législatives dans leurs régions respectives. La BnF a également bénéficié de son implication forte dans l'Observatoire de la web campagne, mis en place par le Forum des droits sur l'internet à l'occasion de cet événement.

nous avons ainsi mis en ligne sur les interfaces de consultation de la Bibliothèque un « parcours guidé » intitulé « Cliquez, votez : l'Internet électoral ».

Ce produit documentaire s'apparente à un choix éditorial d'une vingtaine de thèmes et parcours de navigation tels que « La caricature politique » ou « La stratégie de propagande des principaux candidats », illustrés par des exemples d'archives à partir desquels l'utilisateur peut cliquer pour naviguer ensuite dans la collection. Un prochain parcours sera proposé à l'automne 2008, qui portera sur les écritures du web, notamment l'évolution du genre du journal intime décliné sur les blogs [11]. D'autres projets associant sélection et valorisation ont été lancés en 2008, sur le thème des politiques du développement durable (avec le concours du ministère de l'Écologie) et du web militant (en partenariat avec le Centre d'histoire sociale de Paris-I, la Bibliothèque de documentation internationale contemporaine et des sociologues du Laboratoire usages, créativité, ergonomie de France Télécom).

Grâce à la dynamique engagée à la faveur de ces différents projets, l'équipe des correspondants du dépôt légal de l'internet s'est progressivement étoffée, d'une vingtaine d'agents en 2005 à une centaine en 2008. Sollicités pour leur expertise scientifique ou leur intérêt pour l'internet, ils y consacrent un temps partiel (un à deux jours par mois en moyenne). Des ateliers et des cycles de formation ont permis de développer en interne les compétences propres à la gestion des archives du web, mais aussi de formaliser et de documenter les approches retenues dans les différents domaines. Chaque correspondant veille ainsi sur un secteur spécifique du web pour lequel il identifie une sélection de sites auxquels il attribue ensuite des paramètres de collecte : à quelle fréquence ou à quelle date faut-il que le robot vienne visiter le site ? Est-ce tout le site ou seulement une partie qu'il faut alors copier ? Le site présente-t-il des difficultés techniques à signaler auprès de l'équipe de production ? Est-il nécessaire d'assurer un suivi humain de ce site, ou la collecte large suffit-elle à en assurer une capture correcte chaque année ?

Nouvelles typologies, nouvelles approches

Ces décisions impliquent d'apprivoiser les outils de signalement et de recherche sur l'internet, les techniques de publication en ligne, l'architecture du web. Elles s'appuient sur de nouvelles typologies, qui croisent les caractéristiques éditoriales des sites et leurs modalités d'archivage : un site d'actualité ou de presse doit être archivé fréquemment, mais en surface seulement ; le site d'un ministère ou d'une assemblée peut être archivé une à deux fois par an seulement, mais le plus complètement possible. Ces décisions requièrent également une prise de responsabilité quant à l'économie de l'opération, car lancer une collecte sur un très gros site avec des paramètres inadaptés coûte très cher en stockage pour un résultat qui ne se justifie pas. Toutes ces décisions relèvent bien des compétences d'un bibliothécaire, car c'est toute la qualité et la cohérence de la collection qui viendra boucher les « trous » du dépôt légal constitué en masse par le biais des collectes larges qui sont en jeu.

Pour les professionnels, il s'agit finalement de continuer à gérer un budget d'acquisition (qui s'exprime en octets et en nombre d'URL) dans un domaine thématique ou éditorial (qui échappe souvent aux classifications de Dewey) en utilisant des concepts ou des langages nouveaux (liés à la syntaxe des URL notamment). Appréhender la notion de document patrimonial depuis l'environnement qui est propre au web oblige ainsi à reconsidérer bien des catégories structurantes de nos référentiels documentaires : on l'a vu, les divisions liées au territoire (documentation nationale/étrangère), au support physique (imprimés/audiovisuel/multimédia...), aux thématiques issues de l'encyclopédisme (arts/sciences/sciences humaines...), ou au statut des producteurs (éditeurs patentés/autoproduction ; culture savante/culture populaire) ne sont plus vraiment pertinentes sur le web ou nécessitent d'être réinterprétées [5].

Ce qui change, fondamentalement, c'est la nécessité d'appréhender cette collection à un niveau de granularité supérieur et dans une temporalité qui ne permettent plus de garantir l'exhaustivité ni le repérage de chaque document. On devra souvent renoncer à définir ses choix au niveau des unités (sites, fichiers) au profit d'un pilotage au niveau d'ensembles plus conséquents (domaines, périodes, types d'éditeurs, grappes, etc.). Chacun doit s'appliquer à mieux comprendre le fonctionnement technique, social et culturel du web tout en prenant de la distance avec les documents pris isolément afin d'en comprendre les logiques de regroupement et de lien.

C'est ce double mouvement, paradoxal, de rapprochement et de distanciation qui constitue peut-être la principale difficulté dans l'apprentissage du matériau patrimonial du web. Le succès de l'entreprise dépendra précisément de notre capacité collective à inventer cette nouvelle distance vis-à-vis du document, afin que la collection, même massive, conserve le statut de collection, construite conjointement par des machines et des humains.

Quelles perspectives pour la profession ?

Plusieurs de ces évolutions ne sont pas propres à l'archivage du web. On les retrouve par exemple dans la gestion des ressources électroniques. À l'heure de la concentration commerciale et des bouquets électroniques, l'apport du bibliothécaire se déplace de la sélection des titres vers leur valorisation. Celle-ci s'appuie sur des stratégies de fédération et de personnalisation des services, et sur la conception de l'architecture des interfaces et des systèmes d'information. Dans le domaine de la numérisation de masse, comme l'expérimente actuellement la BnF, l'enjeu est là aussi d'assurer la cohérence et la lisibilité des fonds numérisés tout en s'adaptant à des logiques proprement industrielles. Pour ce faire, on doit alléger certaines des exigences scientifiques que les premières bibliothèques numériques plaçaient en amont de la numérisation et intervenir davantage en aval, dans la qualité de la mise en scène et de la mise à disposition des documents.

Cette nouvelle donne tend d'abord à ramener l'intervention documentaire

à un plus haut niveau de représentation de la collection. Elle déplace ensuite le travail de sélection et de valorisation de son origine vers le terme du processus de traitement : on identifie et on choisit des fonds numériques au milieu de la masse disponible plutôt qu'on ne fait de cette sélection un préalable au traitement. Ce glissement de l'intervention professionnelle requiert des compétences empruntées à d'autres secteurs : l'informatique, bien sûr, mais aussi le droit et la gestion (des marchés, des négociations commerciales), l'industrie et le management (afin d'organiser des chaînes de production à grande échelle). C'est, en tout cas, quelques-unes des compétences qui se dessinent à la BnF à l'aune de l'expérience du dépôt légal de l'internet.

Malgré les avancées significatives des dernières années, le projet de la BnF doit à présent relever des défis tout aussi complexes que les précédents.

Le premier est la montée en charge du dispositif de production afin d'assurer ce qu'on pourrait appeler l'indépendance patrimoniale de la BnF dans le domaine de l'internet. C'est en effet la consolidation du dispositif actuel qui lui permettra d'étendre le champ de son activité et d'y adjoindre de nouveaux partenaires, en particulier les bibliothèques pôles associés. Ce travail a des implications techniques mais aussi humaines, puisqu'il faudra continuer de développer les compétences et la formation des personnels.

Le deuxième défi, qui se joue principalement au niveau international, est le développement des logiciels de collecte, indispensables à l'amélioration de la qualité des captures pour les publications complexes et dynamiques du web 2.0, les bases de données, les cartes, le multimédia et l'audiovisuel en ligne.

Le troisième concerne l'accessibilité et le signalement des collections : en accueillant les archinautes, en suscitant de nouvelles curiosités et de nouvelles approches scientifiques, la Bibliothèque espère à la fois contribuer à l'évolution de la recherche et recueillir l'avis des lecteurs comme des éditeurs du web, qui l'aideront à orienter sa stratégie dans ces domaines.

Enfin, le lancement du chantier Spar est l'occasion d'ouvrir un dossier particulièrement stratégique compte tenu des risques encourus : assurer la conservation pérenne des gigantesques collections extraites du web.

Collecte, coopération, formation, signalement, valorisation, préservation… Ne serait-ce pas une bibliothèque qui se construit autour du web ? ●

Septembre 2008

Bibliographie et références en ligne

[1] Afnic, *Observatoire 2007 du marché des noms de domaine en France*, Saint-Quentin-en-Yvelines, Afnic, 2007, 96 p. www.afnic.fr/data/actu/public/2007/afnic-observatoire-domaines-france-2007.pdf (consulté le 8 septembre 2008).

[2] Valérie Game ; Gildas Illien, « Le dépôt légal d'internet à la Bibliothèque nationale de France : cadre juridique, modèle de collecte, évolutions des métiers », *BBF*, 2006, n° 3, p. 82-85. http://bbf.enssib.fr (consulté le 8 septembre 2008).

[3] Valérie Game ; Clément Oury, « Le dépôt légal de l'internet à la BnF : adapter une mission patrimoniale à l'économie de l'immatériel », communication au colloque Patrimoine et économie de l'immatériel organisé par l'Institut national du patrimoine (Paris, France), 2008.
À paraître en ligne sur le site de l'INP : www.inp.fr

[4] IIPC. International Internet Preservation Consortium www.netpreserve.org (consulté le 6 septembre 2008).

[5] Gildas Illien, « Re-inventing Collection Development Policy in the Age of Web Archiving », actes de la 37e conférence de Liber, Ligue des bibliothèques européennes de recherche (Koc, Turquie), 2008.
À paraître en ligne dans *Liber Quarterly* : http://liber.library.uu.nl

[6] Gildas Illien, « Les mémoires de la Toile : enjeux de la coopération internationale pour l'archivage de l'internet. Le consortium IIPC, une chance pour les bibliothèques francophones ? », actes du 1er Congrès de l'AIFBD, Association internationale francophone des bibliothécaires et documentalistes (Montréal, Canada), 2008 [à paraître].

[7] Gildas Illien, « L'archivage d'internet, un défi pour les décideurs et les bibliothécaires : scénarios d'organisation et d'évaluation ; l'expérience du consortium IIPC et de la BnF », actes du 74e Congrès de l'Ifla, Fédération internationale des associations de bibliothécaires et d'institutions (Québec, Canada), 2008. www.ifla.org/IV/ifla74/papers/107-Illien-fr.pdf (consulté le 8 septembre 2008).

[8] Gildas Illien ; Sara Aubry ; Younès Hafri ; France Lasfargues, *Sketching and checking quality for web archives : a first stage report from BnF*, 2006, 35 p. http://bibnum.bnf.fr/conservation/bnf-qualityforwebarchives-feb06.pdf (consulté le 8 septembre 2008).

[9] France Lasfargues ; Clément Oury ; Bert Wendland, « Legal deposit of the French Web : harvesting strategies for a national domain », actes de la 8e Conférence IWAW, International Web Archiving Workshop (Åarhus, Danemark), 2008. http://iwaw.net/08/IWAW2008-Lasfargues.pdf (consulté le 7 septembre 2008).

[10] Julien Masanès (ed.), *Web Archiving*, Berlin, Heidelberg, New York, Springer, 2006, 234 p.

[11] Bernard Massip, « Une collaboration entre l'APA et la BnF : L'archivage des journaux personnels en ligne », *La Faute à Rousseau*, n° 47, 2008. http://sitapa.free.fr/Documents %20word/collab_bnf_apa. pdf (consulté le 3 septembre 2008).