La sélection de sites web dans une bibliothèque nationale encyclopédique

Sylvie Bonnel, Clément Oury

Submitted on:12/26/2014

La sélection de sites web dans une bibliothèque nationale encyclopédique : une politique documentaire partagée pour le dépôt légal de l'internet à la BnF

Sylvie Bonnel
Coordination service for the Collections Direction, Bibliothèque nationale de France, Paris,France.
sylvie.bonnel@bnf.fr

Clément Oury
Legal deposit department, Bibliothèque nationale de France, Paris, France.
clement.oury@bnf.fr

Abstract

En quelques années, le web est devenu l'un des principaux vecteurs d'expression et de consommation culturelles de la société française ; les publications en ligne ont rejoint notre patrimoine. Celui-ci est d'autant plus précieux qu'il est fragile. En France, il a été décidé d'inscrire la mission de conservation de l'internet dans le sillage pluriséculaire du dépôt légal.
Cependant, l'adaptation de ce dispositif juridique et scientifique à un espace de diffusion aussi vaste et étendu n'a rien d'évident. La BnF définit son périmètre de collecte par une série de restrictions successives : juridiques, techniques et économiques. Pour assurer la représentativité de son dépôt légal, la BnF a également adopté un modèle original d'archivage, qui associe des collectes « larges » du domaine national et des approches plus ciblées de sites identifiés par des bibliothécaires de la BnF ou par des partenaires.
La BnF a ainsi été amenée à appliquer des logiques de sélection dans un cadre de dépôt légal. À cette fin, chaque département associé à la collecte du web a élaboré, au fil des expérimentations, sa propre stratégie documentaire. Les « correspondants » du dépôt légal du web ont adopté des logiques non pas contradictoires mais complémentaires : sélection / échantillonnage, continuité des collections / exploration de nouveaux territoires. C'est désormais à une synthèse de ces différentes politiques que la BnF doit s'atteler, dans le cadre de la refonte de sa charte documentaire et dans un contexte où les contraintes budgétaires appellent à la définition de priorités plus affirmées.

Keywords: Politique documentaire dépôt légal internet

Introduction: conserver la mémoire d'un nouveau média

Qu'y a-t-il de commun entre le site web de Jacques Chirac durant la campagne de la présidentielle de 2002, la communauté de lecteurs *Zazieweb*, ou encore le blog *Mauvais Genres*consacré à la littérature policière et à la science-fiction ? Tous, malgré leur intérêt et leur popularité au moment où ils étaient disponibles en ligne, ont aujourd'hui disparu. Ou plutôt, ils auraient disparu s'ils n'avaient pas été archivés par la BnF au titre du dépôt légal.
En quelques années, le web est devenu l'un des principaux vecteurs d'expression et de consommation culturelles de la société française. L'ensemble des types de documents que les bibliothèques ont la mission de conserver et de communiquer ont connu, avec des calendriers parfois différenciés, leur révolution numérique. Le web a également généré ses propres formes d'expression, qui tirent parti des formidables capacités offertes par ce réseau pour publier, pour lier les contenus les uns aux autres, ou encore pour intégrer des types documentaires hétérogènes : c'est ainsi le cas des blogs ou des réseaux sociaux.
Ce dynamisme exceptionnel a cependant son revers : les contenus en ligne, s'ils sont accessibles depuis le monde entier, sont généralement hébergés sur un nombre restreint de serveurs – quand ils ne disposent pas que d'un seul lieu de conservation. Un problème technique – changement de l'architecture d'un site, défaillance du serveur… – ou une décision humaine peuvent définitivement faire disparaître un contenu. Les études de l'AFNIC, l'organisme qui gère le domaine .fr, ont montré que plus de 20 % des noms de domaine ne se renouvellent pas chaque année[1]. Le web d'hier a déjà largement disparu : ses principaux acteurs à la fin des années 1990, comme Geocities ou Altavista, ont sombré. Dans 10 ou 15 ans, comment pourra-t-on se souvenir de ce qu'était le web d'aujourd'hui ?

Le web, à la fois par le nombre et la variété des contenus qu'il met à disposition, et par la place qu'il a prise au cœur des sociétés contemporaines, est ainsi devenu une part majeure de notre patrimoine. Celui-ci est d'autant plus précieux qu'il est fragile et hautement volatil. Dès les lendemains de la naissance du web, le besoin d'en conserver les traces a été reconnu par quelques institutions pionnières : des associations ou des fondations à but non lucratif, comme Internet Archive, ou des bibliothèques nationales comme celle de Suède, ont commencé dès 1996 à expérimenter les principes et les méthodes d'un archivage à des fins patrimoniales.
En France, les premières réflexions datent de la toute fin des années 1990 ; les premières réalisations, du début de ce millénaire. D'emblée, il a été décidé d'inscrire cette mission dans le sillage pluriséculaire du dépôt légal. Instauré en 1537, ce dispositif édicte que toute publication produite ou diffusée en France doit entrer dans les collections nationales. Depuis lors, il s'est adapté aux différentes évolutions du monde éditorial : après les imprimés, les estampes, le son, la vidéo, les logiciels se sont vus inclus dans la typologie documentaire soumise au dépôt. L'une des spécificités du dépôt légal est son caractère non sélectif : toute production culturelle a vocation à être déposée, quelle que soit la « valeur » que les bibliothécaires lui attribuent. Ce principe vaut à la BnF de conserver aujourd'hui des fonds dont aucune autre bibliothèque n'avait voulu à l'époque de leur publication. D'autre part, ce principe s'accorde bien avec la vocation encyclopédique de l'établissement, réaffirmée dans le décret fondateur de la BnF en 1994[2].
Si « tous les champs de la connaissance », selon les termes du décret de 1994, sont bien représentés sur l'internet, l'adaptation du dépôt légal à un espace de diffusion aussi vaste et

[1] http://www.afnic.fr/fr/ressources/publications/observatoire-du-marche-des-noms-de-domaine-en-france/edition-en-ligne-2014/facteurs-determinants-des-taux-de-renouvellement-3.html.
[2] http://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000000545891.

étendu n'avait cependant rien d'évident. Comment définir la mission d'une institution nationale face au caractère éminemment international du web ? Comment faire cohabiter, tout à la fois, l'exigence d'une couverture aussi complète que possible de la production documentaire, qui est celle du dépôt légal ; la volonté de conserver de préférence ses segments les plus riches et les plus dynamiques, qui correspond aux principes de constitution des collections ; et les contraintes techniques et économiques qui rendent impossible toute volonté d'archivage exhaustif du web ?
Face à ces exigences contradictoires, la BnF a dû constituer, en une dizaine d'années, un modèle original de collecte. Ce modèle découle en premier lieu du code du patrimoine. Il tient ensuite compte des contraintes techniques liées aux modalités de l'archivage du web. Enfin, après avoir déterminé ce qu'il doit collecter, puis ce qu'il peut archiver, il procède d'une approche économique pour identifier ce qu'il veut capturer plus spécifiquement. La BnF a ainsi été amenée à définir des logiques de sélection dans un cadre de dépôt légal ; à cette fin, chaque département associé à la collecte du web a élaboré, au fil des hésitations et des expérimentations, sa propre politique documentaire. C'est désormais à une synthèse de ces différentes politiques que la BnF doit s'atteler, dans le cadre de la refonte de sa charte documentaire.

1) La définition d'un périmètre de la collecte

Des restrictions juridiques
Le périmètre du dépôt légal procède bien entendu de la loi. Le dispositif juridique est établi en deux temps : la loi qui étend le dépôt légal à l'internet est votée en 2006 ; son décret d'application publié en 2011 [1]. La loi définit très largement l'objet du dépôt : y sont soumis tous les « signes, signaux, écrits, sons ou messages de toutes natures faisant l'objet d'une communication au public par voie électronique ». Dans l'esprit du dépôt légal, l'ensemble des publications en ligne sont donc visées, qu'il s'agisse de contenus textuels, de vidéos, de jeux en ligne... à condition que les contenus ne ressortent pas de la correspondance privée. Ainsi, les parties privées des réseaux sociaux sont exclues de la collecte, tandis que les pages publiques y sont comprises.
Le décret apporte un certain nombre de précisions[3]. Il définit d'abord ce que l'on doit entendre comme l'internet « français » : il s'agit tout d'abord des sites hébergés sur des « domaines de haut niveau » français (.fr, .paris, .re pour l'île de la Réunion, etc.) ; et/ou des sites dans un nom de domaine enregistré par une personne domiciliée en France ; et/ou enfin des sites produits sur le territoire français.

Par ailleurs, le décret fixe la répartition des tâches entre les deux institutions dépositaires. L'Institut national de l'audiovisuel se voit confier la charge des sites de télévision et de radio, ainsi que les sites qui y sont « principalement consacrés » ; le champ de la BnF est quant à lui défini par défaut : l'ensemble du web français… sauf les sites dépendant du périmètre de l'INA.
Enfin, le décret précise les modalités de collecte : tous les noms de domaines doivent faire l'objet d'une collecte ; cependant la profondeur de collecte n'est pas précisée et l'exhaustivité de la collecte de chaque site n'est pas demandée. Le décret indique également une fréquence minimale d'archivage : une fois par an – laissant la possibilité aux établissements dépositaires de collecter certains à un rythme plus poussé.

[3] http://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006074236&idSectionTA=LEGISCTA000025005272.

Des restrictions techniques

Afin de réaliser sa mission, la BnF – comme de très nombreuses autres institutions dans le monde –utilise la technologie des « robots » : il s'agit en fait de logiciels qui, à partir d'une liste d'adresses URL qui leur sont indiquées, parcourent le web de liens en liens pour découvrir et capturer les contenus qu'on les a chargés de moissonner[4]. Ces robots permettent d'archiver de très grands volumes de données, mais ils connaissent des limitations. Ils couvrent mal le web « profond » : les bases de données, les documents hautement interactifs y échappent. Les vidéos sont également souvent mal récupérées, sauf à mettre en œuvre des campagnes d'archivage successives. Ces limitations constituent une seconde restriction au périmètre de la collecte : l'objet du dépôt légal… est avant tout ce que l'établissement dépositaire sait capturer.

Une approche économique : le modèle intégré

Même en tenant compte de ces deux restrictions successives, le champ à couvrir reste excessivement vaste : le web français est estimé à sept ou huit millions de sites, dont certains ont un taux de renouvellement très rapide[5].

L'exhaustivité n'est donc plus un objectif atteignable, ne serait-ce que pour des questions de coût et de soutenabilité financière. Cet idéal a donc été remplacé par un objectif de représentativité : il s'agit de constituer une image incomplète, mais fidèle, de l'internet français, qui prenne en compte tous les types de publications et tous les contenus, du plus sérieux au plus dérisoire. A cette fin, la BnF conjugue deux modèles de collecte : le premier est la collecte « large », réalisée une fois par an. Elle couvre l'ensemble des sites web que la BnF a pu identifier comme français – soit plus de quatre millions à ce jour. La profondeur de collecte n'est pas très étendue (quelques milliers de fichiers par nom de domaine), mais cela permet de capturer de façon satisfaisante plus de 90% des sites identifiés.

Cette collecte large est complétée par des collectes « ciblées » qui concernent des sites à capturer plus fréquemment (jusqu'à une fois par an) ou plus profondément (jusqu'à plusieurs centaines de milliers de fichiers par domaine) ; il peut également s'agir de ressources à collecter en raison de leur lien à un événement donné (élections, festivals, rencontres sportives…). L'identification de ces sites est réalisée de manière collaborative au sein de la BnF par un réseau d'une centaine de « correspondants DLweb », chargés à la fois de la sélection, du contrôle qualité et de la valorisation des contenus qu'ils ont demandés. Ces correspondants sont répartis dans une douzaine de départements de collections thématiques (Littérature et Arts, Sciences…) ou de dépôt légal (Cartes et Plans, Estampes, Audiovisuel…). Ces agents sont parfois, dans le cadre de projets, épaulés par des bibliothécaires ou des chercheurs partenaires (dans des bibliothèques régionales ou universitaires, dans des laboratoires de recherche, etc.).

Enfin, conformément à sa démarche de soutenabilité financière, la BnF fixe chaque année la taille maximale de la collection qu'elle a les moyens de constituer et de conserver sur le long terme ; en 2014, la volumétrie annuelle à collecter a ainsi été établie à 100 téraoctets (To). Sur ce total, 50% des ressources sont affectées à la collecte large ; les autres 50% aux

[4]La plupart des outils de collecte et d'accès aux archives du web utilisés par la BnF sont des outils *open source* ; ils ont été développés au sein du consortium international pour la préservation de l'internet, ou IIPC. Voir http://www.netpreserve.org/.

[5]L'AFNIC considère que le domaine .fr représente plus du tiers des sites français ; il y a actuellement 2.7 millions de sites en .fr (http://www.afnic.fr/data/actu/public/2010/afnic-french-domain-name-report-2010.pdf, p. 7).

collectes ciblées. Ainsi, les sélectionneurs de la BnF doivent tenir compte d'un « budget », qui n'est plus libellé en euros mais en nombre de fichiers ou en octets.

Pour accompagner les sélectionneurs dans leur travail, un outil a été mis à disposition : « BnF Collecte du web » ou « BCweb ». On peut, pour chaque site, indiquer une adresse URL, des descripteurs, ainsi que des paramètres de collecte. Ceux-ci sont de trois ordres : la fréquence (de quotidienne à annuelle), la profondeur (tout le site, une sous-partie, juste une page…) et le « budget », c'est-à-dire la taille-cible (en nombre de fichiers).
Ces paramètres sont essentiels car ils permettent de définir plus finement l'objet de la collecte. Ils procèdent eux-mêmes d'une approche économique : il faut donc se demander, pour chaque site, si on veut l'archiver très souvent, quitte à se contenter des pages autour de la « Une » (c'est le cas des sites de presse) ; ou très en profondeur mais peu fréquemment.
Cette approche économique vise ainsi à définir des modèles éditoriaux standards et partagés : site officiel ou de recherche, site d'actualité, blog… Cependant, ces modèles ne disent rien sur le contenu même des sites, dont le choix reste à la discrétion des différents départements impliqués dans la sélection. C'est à leur niveau qu'ont été définies, par maturation progressive, des politiques documentaires.

2) Des politiques documentaires pour les collectes ciblées

Des principes fondateurs
Les collectes ciblées visent à combler les principales lacunes de la collecte large. C'est donc à ce type de collecte que s'applique le mieux le concept de politique documentaire, par ce qu'elle implique en termes de sélection et de construction intellectuelle pour venir compléter le dépôt légal constitué en masse par le biais des collectes larges.

Les stratégies documentaires du dépôt légal du web mises en place progressivement par l'établissement s'inscrivent dans le respect de certains grands principes fondateurs de la BnF affirmés dans sa charte documentaire en 2005 :

- La France objet d'étude privilégié : si la collecte large porte sur les domaines enregistrés en France, les collectes ciblées sont également très fortement concentrées sur les sites français. Néanmoins, cette dimension territoriale peut poser question, voire sembler antithétique du web, par essence peu cantonné à des frontières nationales, et certains chargés de collection ont entrepris, de manière marginale, de collecter des sites étrangers complémentaires du domaine français, selon le modèle des acquisitions étrangères de documents sur support.

- L'encyclopédisme : tous les domaines de la connaissance y sont représentés, comme pour l'ensemble des collections de la BnF, que ce soit les langues et littératures, les sciences humaines et sociales, les sciences et techniques, les arts…

- La dimension temporelle : l'objectif est bien d'allier des collectes « au long cours », visant à la complétude des fonds existants sur tous les supports, et des collectes étroitement liées à l'actualité. Ainsi, en 2013, des collectes ont été spécifiquement consacrées à la guerre au Mali, à la loi sur lemariage pour les couples de même sexe, ou à l'élection du Pape.

Les collectes courantes
Depuis que la collecte est internalisée (2006), le réseau des correspondants du dépôt légal du web a élaboré des stratégies documentaires en adoptant des logiques non pas contradictoires

mais complémentaires : sélection / échantillonnage, continuité des collections / exploration de nouveaux territoires.

On peut identifier « deux approches : la sélection et l'échantillonnage. La première option implique une sélection des sites à collecter en amont, habituellement sur la base d'un jugement de la qualité ou de la valeur scientifique ou esthétique du site (…). Cette approche est analogue à l'acquisition d'ouvrages choisis par un bibliothécaire, avec une logique de sélection d'unités qui enrichissent les collections de recherche. L'approche alternative, l'échantillonnage, est proche du principe du dépôt légal : des sites sont collectés sans jugement préalable sur leur « valeur » ou de leur intérêt potentiel pour les chercheurs d'aujourd'hui ou de demain. Le but est plutôt de préserver un échantillon représentatif de la production nationale née numérique, qui présente le « caractère » du web national à un moment donné. » [1].
On peut ainsi schématiquement observer une double logique, celle des départements uniquement acquéreurs où règne une culture de la sélection destinée à élaborer une collection visant à l'excellence tant dans chacune de ses composantes que dans sa structure, et celle des départements gestionnaires du dépôt légal, qui ont l'expérience d'une collecte visant à une certaine exhaustivité[6].

D'autre part, le web est abordé sous deux angles différents : la continuité des collections et l'exploration de nouveaux territoires.
Les collectes courantes visent à rassembler, dans une démarche de prolongement des acquisitions sur supports, une documentation, désormais largement dématérialisée, toujours plus nombreuses et diversifiée : les publications officielles, les partitions musicales, les cartes et les plans, la musique et le cinéma, toutes sortes d'éphémères ou les périodiques. L'internet est aussi un domaine où la publication est plus souple et moins onéreuse que la publication sur supports, rendant ainsi possible la mise à disposition de nouveaux contenus de toutes sortes représentant des sources pour la recherche : littérature grise, disciplines peu représentées dans la production imprimée…
Mais le web est également considéré en tant que tel, comme source et objet d'étude en perpétuelle évolution, avec ses caractéristiques propres : hypertextualité, innovation graphique, interactivité, actualisation en temps réel...

En observant de plus près les collectes ciblées mises en place par les différents départements de la BnF, force est de constater que les stratégies mises en place allient les différentes logiques, afin de « couvrir » l'éventail des possibles offerts par le web.

On peut tout d'abord prendre l'exemple des collectes de sites de presse. D'une part, la logique de continuité des collections joue pleinement dans la volonté de la BnF de capturer, par l'intermédiaire des robots, les versions PDF des éditions locales de titres de presse quotidienne régionale[7]. D'autre part, une attention est portée aux nouvelles formes de la

[6] A la Bibliothèque nationale de France, la collecte et le traitement des documents reçus par dépôt légal sont organisés par type de documents et confiés à cinq départements gestionnaires : le département du Dépôt légal pour les documents imprimés (livres, brochures, périodiques) et les documents numériques en ligne ; le département de l'Audiovisuel pour les documents sonores, audiovisuels, multimédias et tous les documents numériques sur support ; le département des Estampes et de la photographie pour l'estampe, la photographie, les affiches et l'imagerie ; le département de la Musique pour la musique imprimée (partitions) ; le département des Cartes et plans pour les documents cartographiques (monographies, cartes, plans, globes, atlas, etc.).
[7] Pour des raisons de coût, les versions papier des éditions locales de titres de presse quotidienne régionale ne sont plus conservées à la BnF. Sur ce sujet, voir [2].

presse en ligne : *pure players* comme Médiapart ou Rue89, portails qui relaient l'information des sites de presse…

Autre domaine où se retrouve cette tension entre continuité et nouveauté, celui des cartes : plus de 300 sitesfont l'objet de collectes ciblées dans le domaine de la cartographie et de l'itinéraire, depuis les sites institutionnels les plus connus jusqu'aux blogs d'amateurs de cartes. Pour le département des Cartes et Plans de la BnF, il s'agit d'une continuité de ses missions de conservation des contenus physiques, pour conserverune trace, non pas réellement exhaustive mais plutôt représentative, de la manière dont on se représente l'espace à un moment donné, et des usages qui peuvent être retranscrits de manière cartographique [3].

De la même manière, dans le domaine des arts du spectacle, on allie logique d'exhaustivité et sélection : les sites, généralement de taille réduite, sont relativement bien captés par la collecte large, ce qui permet un échantillonnage significatif de ce qui est produit sur le web français. Les collectes ciblées ont donc pour objectif d'enrichir ces archives. La sélection des sites s'est construite par strates successives : d'abord sur les sites institutionnels de théâtres peu documentés, afin de compléter le fonds documentaire par le web, puis également sur les autres spécialités du département. Une attention particulière a également été portée à la critique théâtrale, grâce à une sélection de blogs. Des sites d'information sans lien avec une structure théâtrale, des sites mettant en ligne des sources et, en général, des sites qui n'étaient pas la transposition ou la continuité de ce qui se produisait sur un autre support, sont venus enrichir la sélection. Enfin, en 2013, un nouvel ensemble a été ajouté, une douzaine de sites web en lien avec des fonds d'archives conservés au département des Arts du spectacle. Les arts du spectacle représentent, à l'heure actuelle, plus de 180 sitessélectionnés [4].

Dernier exemple, enfin : les collectes courantes concernant l'audiovisuel ontété conçues autour de trois axes majeurs : la continuité de la collecte par rapport aux supports audiovisuels et multimédias pour archiver les contenus sonores, audiovisuels et multimédia diffusés sur l'internet, dont les producteurs ou les éditeurs étaient traditionnellement (ou sont encore en parallèle) déposants au dépôt légal des phonogrammes, vidéogrammes ou documents multimédia sur supports. Par extension, ce principe de continuité s'applique à des sites diffusant des contenus sonores, audiovisuels ou multimédia qui auraient pu auparavant être diffusés sur support, mais dont les producteurs ou éditeurs ont directement commencé avec l'internet. Le deuxième axe adopté consiste à collecter les sites qui représentent des formes novatrices de création et/ou de diffusion sonore, audiovisuelle ou multimédia apparues avec l'internet, tels que des sites de partage de contenus générés par les utilisateurs, des sites de Net Art… Enfin, la collecte vient, en prolongement d'acquisitions sur supports, collecter les sites documentant le monde de la musique et de la création sonore, du cinéma et de l'audiovisuel, du jeu vidéo et de la création numérique. L'audiovisuel, en alliant volonté d'exhaustivité, logique de référence et logique de représentativité, représente actuellement près de 5 000 sitessélectionnés [5].

Les collectes projets

Aux collectes courantes, viennent s'ajouter les collectes projets. Elles ont un périmètre plus restreint que les collectes courantes, à la fois au niveau thématique mais également dans le temps. Elles répondent à un besoin documentaire identifié, au périmètre délimité, et n'ont, par principe, pas vocation à devenir permanentes. Elles sont généralement le fruit de coopérations, à la fois internes entre départements de la BnF, mais également avec des partenaires extérieurs.

La première collecte projet, mise en place sous forme expérimentale dès 2002 puis systématisée après la loi de 2006, a porté sur les campagnes électorales[6]. Ces dernières, en effet, se jouent désormais aussi bien sur le web que dans la rue. Les sites des responsables politiques ou des militants constituent un matériau précieux pour comprendre les enjeux et l'issue d'un scrutin, mais ils sont par essence extrêmement volatils. La BnF a donc mis en place un dispositif de sélection, d'archivage et d'accès à ce nouveau type de sources, et l'a appliqué à la plupart des élections depuis 2002, nationales ou locales. Ces collectes ont notamment permis de documenter l'usage des nouvelles technologies par les partis : ainsi, les blogs politiques dominent la campagne en 2007 ; l'usage de réseaux sociaux comme Facebook ou Twitter est timide en 2009 puis s'affirme à partir de 2010.

Un autre exemple de collecte ciblée est celle touchant portant sur les journaux intimes, qui, avec la mode des blogs, ont connu un regain considérable. Pour préserver « le champ inédit et immense de l'expression autobiographique qui s'y est constitué », l'archivage du web est devenu une nécessité, assurée à la fois par l'équipe du département Littérature et Art de la BnF pour les blogs et les sites d'écrivains, et par un partenaire, l'Association pour l'autobiographie et le patrimoine autobiographique,pour les sites d'expression personnelle [7].

De la sélection à la mise à disposition

L'articulation entre les différents types de collectes, associant des collectes de grande ampleur et des sélections humaines plus ciblées, permet de répondre aux attentes exprimées par les publics lors de l'étude prospective sur les représentations et les attentes des utilisateurs potentiels sur les archives de l'internet, menée par la BnF en 2010-2011. Si les professionnels et le « tout venant » de la bibliothèque de Recherche exprimaient un besoin relatif, souvent ponctuel, de l'archivage du web, les chercheurs, pour leur part, percevant à la fois l'immense richesse et la volatilité de l'univers du web, exprimaient la difficulté de définir et circonscrire des corpus significatifs. Face à cette difficulté, la BnF était clairement perçue comme un tiers de confiance capable de garantir au chercheur l'accès à des collections raisonnées et documentées. En termes de contenus, les attentes des chercheurs indiquaient plusieurs pistes : développer des collectes qui permettent de garder la trace des nœuds importants et des réseaux ; archiver les sites les plus populaires, ainsi que ceux qui font vraiment rupture ; enfin, ne pas collecter seulement des éléments discrets ou isolés mais garder la trace des pratiques du web qui marquent « l'air du temps » et documentent les tendances, sociales, commerciales, etc., à grande échelle. En tout état de cause, pour toutes les catégories interrogées, il ressortait que le concept de sélection par les bibliothécaires était perçu comme légitime, inévitable et nécessaire compte tenu des volumes à archiver [8].

Les attentes des publics portaient également sur une mise à disposition facilitée de la collection ainsi constituée : si l'accès est de par la loi restreint à l'enceinte des niveaux « recherche » de la BnF, l'effort porte sur la proposition d'outils d'accès efficaces, prenant en compte la dimension hypertextuelle de l'internet. La *Wayback Machine*, ou machine à remonter le temps, principale interface de consultation des collections de la BnF, permet aux chercheurs de naviguer dans l'archive du web comme ils auraient navigué à l'époque sur le web vivant. Cette navigation dans l'espace se double d'une exploration diachronique. Partant d'un site donné, on peut remonter le temps et analyser ses transformations successives.
Cette indexation suppose, pour découvrir un site, que l'on connaisse déjà son adresse… Pour pallier à ce défaut, il semble indispensable de mettre en place une indexation « plein-texte », qui permette d'identifier des pages et des fichiers en fonction de leur contenu textuel, à la suite d'une requête par mot-clef. En raison de la très grande volumétrie des collections

(21 milliards de fichiers, 470 To de données), et de la difficulté de gérer des couches temporelles successives, cette indexation plein-texte n'a pour l'instant été réalisée que de façon très partielle, ce qui représente un frein à l'usage des collections.
En parallèle, afin de pallier la difficulté pour des non-initiés de distinguer l'offre des archives de l'internet de l'offre accessible directement sur le web, la BnF met en œuvre une valorisation sous forme de « Parcours guidés » conçus comme des produits d'appel, sur des corpus facilement appréhendables et représentatifs de la mémoire nationale, politique ou culturelle, où le phénomène de la disparition apparaît clairement.

3) Et après ?

Une intégration dans la charte documentaire
La BnF a entrepris d'actualiser sa charte documentaire qui date de 2005. Pour l'actualisation de la charte, dont l'achèvement est prévu pour la fin 2014, plusieurs choix méthodologiques ont été effectués : aborder la politique documentaire dans son ensemble, c'est-à-dire la politique d'enrichissement des collections quel que soit leur mode d'entrée et leur support ; privilégier une présentation par grands domaines disciplinaires, beaucoup plus lisible pour le public, plutôt que par le biais de l'organisation en départements. La charte actualisée comportera une première partie consacrée à un rappel du contexte et à une synthèse des évolutions majeures depuis 20 ans et des perspectives jusqu'en 2020. Le cœur de la charte sera constitué de fiches destinées à donner, par grands domaines disciplinaires, les priorités stratégiques de développement des collections de la BnF.

La charte documentaire des acquisitions de la BnF, publiée en 2005, ne prenait pas en compte le dépôt légal du web. Son actualisation, actuellement en cours, l'inclut : en raison de la place qu'il a prise dans l'activité des chargés de collections, en raison du volume des données entrées, en raison de ses coûts de fonctionnement, en raison surtout de son rôle incontournable pour assurer la complétude des collections patrimoniales, et donc les missions de l'établissement, telles qu'elles sont décrites dans son décret fondateur. Le dépôt légal du web y fera l'objet d'un focus particulier dans la partie liminaire, afin de délimiter les problématiques qu'il pose, puis sera présent dans chaque fiche-domaine thématique puisque la politique documentaire de la BnF est désormais conçue dans une stratégie globale de constitution des collections, qu'elles soient matérielles ou numériques.

Vers l'affirmation de priorités documentaires
Le dépôt légal du web, à l'échelle de la vie de la BnF, s'inscrit dans son histoire très récente. Il a été mis en place à partir d'un schéma technique (le « modèle intégré ») et organisationnel (le réseau des correspondants) raisonné et solide, et s'appuie sur des principes et outils partagés structurants.
Pour les différents acteurs internes, cela a nécessité un temps de formation et d'acculturation, à des logiques, typologies et contraintes techniques auxquelles les chargés de collections n'étaient pas accoutumés, mais il est désormais entendu que les différentes décisions prises par un correspondant du dépôt légal du web dans son travail de sélection relèvent bien des compétences d'un bibliothécaire, en ce qu'elles viennent garantir la qualité, la représentativité et la pertinence des ressources ainsi collectées.
Néanmoins, plusieurs constats s'imposent : outre queles forces humaines qui y sont mises par les différents départements sont très inégales, les notions de coûts, comme pour les documents sur supports, sont hétérogènes selon les domaines : de manière simple, une vidéo « coûte » plus cher (en octets) qu'un site essentiellement textuel ; enfin, les collectes mises en

place par les différents partenaires ne « pèsent » pas le même nombre d'octets selon qu'on est majoritairement dans une logique d'exhaustivité ou de sélection.

Depuis que les collectes ciblées sont internalisées, le choix a été délibéré de ne pas fixer a priori de volumétrie par thématique ou département. Or l'archivage du web doit fonctionner dans une nécessaire maîtrise des coûts de fonctionnement, le risque étant de collecter toujours plus de sites mais toujours plus en surface, de privilégier la quantité aux dépens de la profondeur et donc de la qualité de la collecte.
La politique documentaire inclut traditionnellement dans sa réflexion la notion d'élimination des collections. Or, dans le cas d'espèce, ce qui est collecté dans le cadre du dépôt légal du web est, par définition, destiné à être conservé, à « faire patrimoine ». Maisle réseau des correspondants a une mission de veille constante afin d'ajuster les collectes au plus juste : diminuer la fréquence de collecte pour des sites devenus moins actifs, arrêter la collecte d'un site qui ne semblerait plus pertinent, ajouter à la collectes des sites émergents, … On parlera donc non pas de désherbage a posteriori, mais d'une élimination en creux, a priori, pour le cas de l'archivage de l'internet.
Si jusque-là la volumétrie annuelle allouée (en To) permettait d'assurer l'ensemble des collectes ciblées telles que souhaitées par les correspondants, 2013a été la première année où il est devenu impératif de faire des choix exclusifs afin d'assurer une collecte aussi variée et représentative que possible. Ainsi, il a été décidé d'arrêter les collectes projets considérées comme menées à bien, de réduire le rythme des collectes ciblées dès lors que la vitalité plus ou moins grande des sites collectés le permet, de mener un chantier de dédoublonnage aussi entre les collectes ciblées menées par les différents acteurs, ou encore de contingenter la collecte Vidéo à 10 To annuels.

Après un temps de maturation technique, organisationnelle et intellectuelle, il apparaît désormais nécessaire d'affirmer collectivement des priorités documentaires : comme souvent, la préoccupation de la politique documentaire vient initialement de contraintes budgétaires (ici, de capacité d'archivage en To), et force à se poser les questions en termes de stratégie de long terme, en s'interrogeant sur les collections ayant la plus forte plus-value pour la recherche future.
Pour autant, comment identifier aujourd'hui ce qui intéressera les chercheurs de demain ? La récente conférence d'ouverture du consortium IIPC, réuni à Paris en mai 2014, l'a bien démontré au travers des interventions de plusieurs chercheurs : les champs de recherche sont nombreux, chaque projet mobilisant des constitutions de corpus et outils d'analyse différents, et l'inventivité des chercheurs est aussi grande que la richesse du web. Il est donc sans douteprimordial de « ménager l'avenir » en maintenant la combinaison des différents types de collectes de l'internet au service d'un encyclopédisme raisonné.

Références

[1] Illien Gildas, Sanz Pascal, Sepetjan Sophie, Stirling Peter, « La situation du dépôt légal de l'internet en France : retour sur cette nouvelle législation, sur sa mise en pratique depuis cinq ans, et perspectives pour le futur », *actes du 77e congrès de la Fédération internationale des associations de bibliothécaires et d'institutions (IFLA)*, San Juan (Porto Rico), août 2011. [http://conference.ifla.org/past-wlic/2011/193-stirling-fr.pdf, consulté le 7 juin 2014].

[2] Oury Clément, « When press is not printed*:* the challenge of collecting digital newspapers at the Bibliothèque nationale de France », *actes de la préconférence de la section presse de l'IFLA*, Mikkeli (Finlande), août 2012.[http://www.ifla2012mikkeli.com/getfile.php?file=154ou http://halshs.archives-ouvertes.fr/docs/00/76/90/84/PDF/LegalDepositNewspapersBnF_Oury_IFLA2012.pdf, consulté le 7 juin 2014].

[3] Lebailly Guillaume, Marchand Emmanuelle, « Cartes et carnets de voyage en ligne : la BnF collecte le Web », *La géographie*, 2013, n°1549, p. 42-43.

[4] Obligi Cécile, « Les archives du web à la BnF : un formidable gisement à venir exploiter ! », à paraître dans*Skén&graphie*, n°2, Annales littéraires/Presses Universitaires de Franche-Comté, Besançon, 2014.

[5] Carou Alain, « Archiver la vidéo sur le web », *Bulletin des bibliothèques de France*, n° 2, 2007. [http://bbf.enssib.fr/consulter/bbf-2007-02-0056-012, consulté le 7 juin 2014].

[6] Oury Clément, « Soixante millions de fichiers pour un scrutin. Les collections de sites politiques à la BnF », *Revue de la BnF*, 2012/1 n° 40, p. 84-90. [http://www.cairn.info/resume.php?ID_ARTICLE=RBNF_040_0084, consulté le 7 juin 2014].

[7] Genin Christine, « Collecter l'océan ? L'archivage de l'intime en ligne à la BnF », *Bibliothèque(s)*, n°47-48, décembre 2009, p. 50-52.

[8] Chevallier Philippe, IllienGildas, Stirling Peter, « Web Archives for Researchers: Representations, Expectations and Potential Uses », *D-Lib Magazine*, 2012, vol. 18, no 3/4. [http://www.dlib.org/dlib/march12/stirling/03stirling.html, consulté le 7 juin 2014].