Constituer un réseau d'accès aux archives de l'internet : l'exemple français

Ange Aniesa, Ariane Bouchard

▶ **To cite this version:**

Ange Aniesa, Ariane Bouchard. Constituer un réseau d'accès aux archives de l'internet : l'exemple français. IFLA WLIC 2017, IFLA, Aug 2017, Wroclaw, Pologne. hal-01703421

HAL Id: hal-01703421
https://bnf.hal.science/hal-01703421

Submitted on 7 Feb 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

IFLA WLIC 2017 WROCŁAW

Submitted on: 20.06.2017

Constituer un réseau d'accès aux archives de l'internet : l'exemple français

Ange Aniesa
département du Dépôt légal, Bibliothèque nationale de France, Paris, France
ange.aniesa@bnf.fr

Ariane Bouchard
département du Dépôt légal, Bibliothèque nationale de France, Paris, France
ariane.bouchard@bnf.fr

Copyright © 2017 by Ange Aniesa et Ariane Bouchard. This work is made available under the terms of the Creative Commons Attribution 4.0 International License:
http://creativecommons.org/licenses/by/4.0

Résumé :

Depuis 2006, la BnF a pour mission de collecter l'internet français au titre du dépôt légal. Pour remplir cette mission au mieux, elle a progressivement mis en place un système d'archivage complet et ainsi collecté des milliards de pages web. Sur la base du décret d'application de la loi DADVSI, la BnF a cherché à rendre ses collections d'archives de l'internet, à l'origine uniquement consultables dans ses espaces Recherche, accessibles dans d'autres établissements en région. Cet article présente les différentes étapes de l'ouverture de ces accès : l'habilitation des bibliothèques de dépôt légal imprimeur ; les problématiques organisationnelles et techniques rencontrées et les solutions adoptées ; les enjeux au stade actuel du projet, alors que seize établissements sont déjà équipés d'un service d'accès aux archives de l'internet.

Mots-clés : dépôt légal numérique, accès, archives de l'internet

Engagée dans des expérimentations depuis 2002 et confortée par la publication de la loi DADVSI instaurant le dépôt légal de l'internet en 2006 [1], la BnF a progressivement mis en place un cycle d'archivage du web. Ce circuit est achevé entre 2011 et 2013, avec deux mises en production : celle d'un outil de sélection, BnF Collecte du web (BCweb)[1], et celle d'une filière de conservation des archives de l'internet dans le système de préservation numérique de la bibliothèque (SPAR). La fonction d'accès était déjà assurée depuis 2008 grâce au logiciel Wayback Machine accompagné de briques techniques développées en interne.

Sur ces bases solides, la BnF s'est engagée en faveur de l'accès aux archives de l'internet hors de ses salles de recherche, dans lesquelles la consultation est réservée à des chercheurs dûment accrédités. Cette volonté s'est affichée dès 2009 lors d'une journée réunissant la BnF et les bibliothèques de dépôt légal imprimeur (BDLI) : l'obtention d'un cadre juridique autorisant l'ouverture de nouveaux accès est alors citée comme une condition

[1] Sur le processus de sélection, voir [2]

nécessaire à la coopération nationale autour du dépôt légal de l'internet. Le décret d'application de la loi DADVSI, promulgué en 2011, prévoit cette possibilité, même s'il n'en précise pas les modalités : l'article R132-23-2 du Code du patrimoine spécifie que la consultation des archives s'effectue « à la Bibliothèque nationale de France et dans tout organisme habilité à mettre en œuvre cette consultation par arrêté du ministère de la culture »[2]. Il restait donc à définir la liste des organismes susceptibles d'être habilités. Pour ce faire, l'Institut national de l'audiovisuel (Ina), la BnF et le ministère de la Culture ont poursuivi des discussions qui ont abouti trois ans après la parution du décret.

Le choix des bibliothèques de dépôt légal imprimeur

L'arrêté est en effet pris par le ministère de la Culture le 16 septembre 2014. Il habilite 26 bibliothèques de dépôt légal imprimeur à accueillir un accès aux archives de l'internet.

Cette décision repose sur des arguments anciens et solides. Les établissements qualifiés comme BDLI – le plus souvent des bibliothèques, plus rarement des archives départementales – assurent, outre les missions classiques, la collecte, la conservation et la valorisation des documents reçus par le biais du dépôt légal imprimeur (DLI)[3]. Cette mission de dépôt légal, complémentaire du dépôt légal éditeur, vise à constituer des fonds régionaux. 29 établissements sont habilités pour cette mission, ce qui correspond à peu près au nombre des anciennes régions administratives françaises[4]. Sur les 29, 26 ont signé une convention avec la Bibliothèque nationale de France leur conférant le statut de pôles associés du dépôt légal[5]. Dans ce contexte, le dépôt légal de l'internet peut être considéré comme une forme d'extension de la mission du dépôt légal imprimeur assurée par ces établissements à rayonnement régional, de sorte que, d'après l'arrêté de 2014, un simple avenant à la convention de coopération passée avec la BnF leur suffit pour demander l'ouverture de la consultation des archives de l'internet.

La plupart de ces établissements étaient en outre déjà impliqués dans l'archivage du web à l'époque de la parution de l'arrêté – certaines depuis près de 10 ans – par le biais des collectes électorales. En 2004, 4 bibliothèques se sont portées volontaires pour participer au processus de sélection des sites, elles seront 8 pour les législatives de 2007, 19 en 2010 et 20 en 2012. Lors de cette dernière collecte, les sélectionneurs ont commencé à saisir directement les adresses des sites dans l'application en ligne BCweb, dès lors qu'ils y étaient habilités. L'ouverture de BCweb a marqué un tournant dans l'implication des BDLI, mieux intégrées dans le processus de sélection.

Les bibliothèques qui ont ainsi pris part à la constitution des collections l'ont logiquement fait valoir. Si l'argument était légitime, il n'était sans doute pas suffisant, puisque la possibilité d'ouvrir la consultation à d'autres organismes participant à la sélection, comme des bibliothèques universitaires, n'a pas été retenue dans l'arrêté de 2014. L'habitude

---

[2] *Code du patrimoine*, article R132-23-2. https://www.legifrance.gouv.fr/affichCodeArticle.do;jsessionid=B3C3AD0534BA50741D457413651888E3.tpdila21v_2?idArticle=LEGIARTI000025004804&cidTexte=LEGITEXT000006074236&dateTexte=20170614 (consulté le 13 juin 2017).

[3] Liste des établissements sur le site www.bnf.fr : http://www.bnf.fr/fr/professionnels/dl_comment_declarer/s.dl_formulaires_bdli.html?first_Art=non (consulté le 13 juin 2017)

[4] Un redécoupage des régions a été opéré par la loi nº 2015-29 du 16 janvier 2015. Le nombre de régions est passé de 22 à 13. Cette loi n'a pas directement touché les BDLI, qui conservent la même couverture territoriale qu'auparavant.

[5] Sur les Pôles associés de dépôt légal imprimeur, voir http://www.bnf.fr/fr/professionnels/cooperation_nationale/a.poles_associes_depot_legal.html (consulté le 13 juin 2017).

de gérer des collections patrimoniales du dépôt légal, le maillage qui garantit une bonne couverture territoriale, ainsi que la tradition de coopération étroite qui les lie à la Bibliothèque nationale de France ont certainement été les principales raisons de l'habilitation des bibliothèques de dépôt légal imprimeur.

Le déploiement de l'accès distant

Anticipant la promulgation du décret, la Bibliothèque nationale de France s'est engagée dès 2012 dans une réflexion sur les modalités de déploiement de l'accès distant aux archives de l'internet. Ce chantier technique s'est déroulé parallèlement au chantier juridique et a abouti à une première ouverture de l'accès distant le même mois que la parution de l'arrêté, en septembre 2014. En effet, l'ouverture des services d'accès posait de nombreuses questions organisationnelles et techniques du fait des particularités des conditions de consultation (accréditation obligatoire, impossibilité de téléchargement…).

Pour y répondre, un groupe pilote constitué de cinq bibliothèques de dépôt légal imprimeur s'est réuni à l'initiative de la BnF en juillet 2012. Il s'agissait de réfléchir à des solutions génériques qui pourraient ensuite s'appliquer à toutes les BDLI, voire à d'autres établissements, dans la perspective d'un éventuel élargissement du champ de l'arrêté. Lors de cette réunion, un certain nombre de questions primordiales ont été soumises au groupe : où installer l'accès ? comment le contrôler ? sur quels postes informatiques ? quels agents pour renseigner le public ? Les réponses étaient loin d'être uniformes, malgré la taille restreinte du groupe. En ce qui concerne le choix de la salle de consultation par exemple, deux options s'offraient généralement : la salle des fonds anciens ou la salle du dépôt légal imprimeur. Selon l'établissement, l'une ou l'autre solution semblait préférable, en fonction de la fréquentation dont ces salles disposaient déjà. La présence ou la perspective d'accueillir les archives de l'internet de l'Ina, institution avec laquelle la BnF partage la mission de dépôt légal du web, pouvait également peser dans la balance. En outre, si l'accès à ces salles était contrôlé ou a minima soumis à inscription, il existait rarement un processus d'accréditation identique à celui pratiqué à la BnF. Pour compléter les premiers éléments rassemblés par le groupe pilote, la BnF a envoyé un questionnaire détaillé à toutes les bibliothèques de dépôt légal imprimeur. Les mêmes sujets y étaient abordés : le lieu d'installation de l'accès, les conditions d'accréditation, le système informatique, le nombre de postes requis. A partir des conclusions tirées de l'enquête, le choix de la salle a été laissé à l'appréciation des bibliothèques. Le processus d'accréditation a été adapté à chaque cas, mais toujours dans l'objectif d'être similaire à celui de la BnF. La solution informatique qui a été choisie est en revanche générique afin qu'elle puisse s'appliquer à l'ensemble des établissements concernés.

A l'issue de cette enquête la Bibliothèque nationale de France a donc opté pour un dispositif informatique souple : le poste qui donne accès aux archives appartient à l'établissement proposant le service de consultation et n'a pas besoin d'être dédié à cette seule consultation. L'accès sur ce même poste à l'internet, à des outils bureautiques et à d'autres ressources numériques proposées par la bibliothèque est même jugé préférable pour permettre une exploitation optimale des archives de l'internet. Le poste doit également être équipé de hauts parleurs ou d'un casque audio pour permettre la lecture de tous les types de contenus. Ce dispositif sécurisé nécessite au préalable d'enregistrer puis d'authentifier le poste devant se connecter au système d'information de la BnF. Le système d'authentification repose sur une solution propriétaire. Il doit être installé une unique fois par poste et s'appuie sur l'attribution d'une licence d'accès par l'administrateur du côté de la BnF et sur

l'installation d'un plug-in sur le poste authentifié du côté la bibliothèque partenaire. Une fois le poste enregistré, les usagers de la bibliothèque utilisent par la suite un navigateur récent, compatible avec les technologies HTML5 et WebSockets, et configuré pour se rendre sur le service via une adresse URL dédiée.

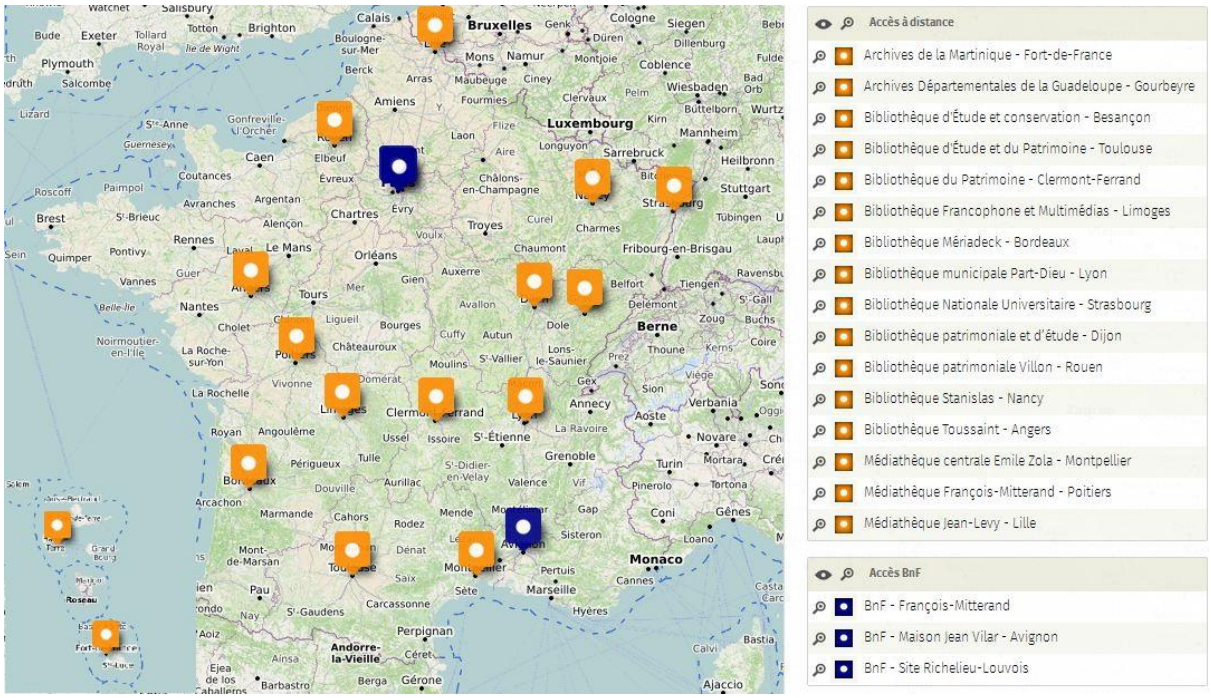
Depuis cette URL, après acceptation de la charte de bon usage des services à distance de la BnF, la session de navigation s'ouvre et donne accès aux collections au moyen d'un environnement sécurisé qui offre des conditions de consultation identiques à celles proposées dans les salles de lecture de la BnF. Un second navigateur de couleur orange encapsulé dans le premier navigateur, ainsi qu'un ensemble de programmes associés permettent d'afficher les contenus web tels qu'ils étaient mis à disposition du public au moment de la collecte ; la recherche, la navigation et la visualisation des contenus archivés sont assurées par le logiciel libre OpenWayback. Conformément à la loi, il est interdit de télécharger les fichiers archivés (textes, images, vidéos…) et le dispositif informatique empêche le téléchargement sur le poste local. En revanche, sous réserve des règles et des outils mis en place dans chaque bibliothèque, il est possible de copier des extraits de texte ou faire des copies d'écran à des fins d'utilisation privée ou de citation. Le chantier de mise en production de cette solution technique a duré du dernier semestre 2013 à juin 2014, tests compris. Lors des tests avec trois bibliothèques pilotes, le dispositif a été jugé simple et rapide à installer. En effet, il nécessite seulement l'envoi des prérequis techniques et un accompagnement par téléphone lors de l'étape d'authentification menée par les services informatiques des deux établissements.

1. Page d'accès à distance aux archives de l'internet

Les trois premiers services de consultation à distance ont été logiquement déployés, en 2014 et début 2015, dans les trois bibliothèques ayant effectué ces tests : la médiathèque centrale Emile Zola de Montpellier, la bibliothèque nationale et universitaire de Strasbourg et la bibliothèque Stanislas de Nancy. Ces trois premières ouvertures ont été l'occasion de perfectionner le dispositif de formation proposé aux équipes de la bibliothèque accueillant ce service. Le principe d'une journée de formation, assurée sur place par les équipes des services du dépôt légal numérique et des systèmes d'information de la BnF a été retenu. Il permet de former les agents médiateurs à une ressource relativement méconnue. En général, la

formation s'étend sur une journée pleine, avec un découpage en deux temps. Le matin est consacré à une présentation générale des archives de l'internet du point de vue documentaire (constitution, nature et contenu des collections) et technique (interface de consultation), s'adressant à toute personne intéressée. L'après-midi est consacré à des ateliers pratiques de présentation et de consultation des archives, à destination des médiateurs prévus sur la ressource et amenés à renseigner les publics. Le fait que la formation soit dispensée dans un premier temps par les agents de la BnF sur place est apparu important et nécessaire aux établissements pour la mobilisation et l'intérêt des équipes. Ils effectuent par la suite un rappel en interne, au moins tous les ans, à leurs agents. Le déroulé a ainsi été appliqué dans les treize établissements qui se sont dotés d'un accès de septembre 2015 à mai 2016. Il a parfois été adapté ou étendu avec une seconde journée de présentation en direction du public.

Durant ces deux ans et demi, la mise en place des accès distants au sein de 16 établissements parmi les 26 établissements éligibles aux archives de l'internet a été échelonnée dans le temps. Le calendrier de déploiement s'est établi progressivement et en concertation avec les partenaires. Il s'est principalement appuyé sur les volontariats exprimés (lors la participation dans le groupe pilote ou à l'occasion de journées professionnelles par exemple). La BnF a également veillé à ce qu'il y ait une bonne répartition géographique des accès, en métropole et dans les départements et région d'outre-Mer, afin d'assurer un accès relativement proche à des chercheurs éloignés de Paris, où ces collections étaient initialement exclusivement disponibles.



2. Où consulter les archives de l'internet de la BnF ?

Les accès aujourd'hui et demain : enjeux et perspectives

Après cette première phase de déploiement, plusieurs enjeux demeurent. Le premier réside paradoxalement dans le maintien d'une solution informatique pérenne pour assurer les accès à long terme. L'obsolescence de la solution technique permettant le système

d'authentification d'un poste fixe oblige actuellement la BnF à trouver une alternative pour équiper les établissements intéressés à l'avenir. La seconde phase de déploiement est conditionnée par la mise en place d'une solution d'authentification sur place tout aussi sécurisée, alors que le marché de l'authentification en informatique est majoritairement tourné vers la mobilité. Cependant, aucun risque d'interruption de service ne pèse sur les accès déjà déployés.

Pour les établissements proposant l'accès, l'enjeu principal dans les mois et années à venir est la valorisation de cette ressource patrimoniale numérique encore méconnue ou peu utilisée par les publics universitaires ou professionnels.

La sélection partagée de sites internet dans le cadre du dépôt légal numérique rejoint cet objectif. En effet, les corpus qui présentent un intérêt régional sont les plus immédiatement valorisables et exploitables pour ces établissements. Or, en plus de bénéficier de 31 titres de presse locale collectés dans leur version numérique payante [3], soit 260 éditions différentes[6], les bibliothèques participent de manière accrue à la constitution de collections d'archives de l'internet à l'échelle locale. Cela alimente les archives de l'internet avec des corpus régionaux de plus en plus complets, car finement sélectionnés par des professionnels sur place, spécialistes de ce type de ressources, contribuant ainsi à l'ancrage de ces collections dans le paysage documentaire régional. Ces sélections thématiques de centaines de sites relatifs à une région, suivant une politique documentaire préétablie et commune à tous les établissements, s'organisant autour d'une quinzaine de thème. A ce jour, trois collectes web régionales ont été créées. Les collectes électorales sont également un bon point d'entrée pour découvrir des captures d'archives de l'internet : directement constituées par les établissements, elles représentent un enjeu fort pour la documentation de la vie politique locale, les manifestations de l'exercice de la citoyenneté et la sauvegarde des traces des campagnes électorales, dont la durée de vie sur le web vivant est de plus en plus réduite. La sélection partagée contribue indirectement à cet objectif.

Les stratégies de valorisation au sein des bibliothèques proposant l'accès passent assez classiquement une bonne signalétique, accompagnée d'une communication traditionnelle, à partir de supports préexistants (dépliants et kakémonos) mis à disposition par la BnF. L'aménagement d'une salle ou d'un espace dédié au patrimoine numérique regroupant les collections de la BnF et celles spécifiques aux archives du dépôt légal web de l'Ina permet de réunir ces collections complémentaires, et ainsi de renforcer l'attractivité du service. La médiation en ligne, via les blogs ou les réseaux sociaux et directement vers les publics potentiels, est aussi privilégiée. Elle est néanmoins par définition limitée par les règles de réutilisation des contenus des archives de l'internet, dans le cadre du droit d'auteur et du droit à la vie privée. Une prospection active des publics potentiels est ensuite effectuée sous plusieurs formes par les établissements : par l'accompagnement des usagers sur place, avec des présentations spécifiques et programmées ; par l'organisation de journées d'étude et/ou de présentations de l'activité de dépôt légal imprimeur, au sein de laquelle s'inscrivent les archives de l'internet ; par le rapprochement avec des lieux axés sur le numérique (fablab, écoles spécialisées, « fabriques » du numérique) ou des actions de communication vers l'ensemble des agents des sites universitaires régionaux ; par la coopération avec des laboratoires de recherche en informatique et information scientifique et technique, avec les maisons des Sciences de l'Homme ; enfin par des prises de contact avec l'association des professionnels des bibliothèques (ABF), les unités régionales de formation à l'information

[6] Chiffre au 1er mai 2017

scientifique et technique (URFIST) ou encore l'École nationale supérieure des sciences de l'information et des bibliothèques (ENSSIB). Les évolutions à venir des fonctionnalités de recherche dans les archives de l'internet (indexation plein texte des collections, constitution de corpus) devraient à terme renforcer l'attractivité des collections en les rendant plus faciles à explorer et à exploiter.

Ainsi, la constitution partagée des collections d'archives de l'internet et leur utilisation sont désormais des enjeux communs.

Références

[1] Illien G., Sanz P., Sepetjan S., Stirling P. 2012. The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future. In *IFLA journal,* 2012, vol. 38, n° 1. [http://www.ifla.org/files/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf]

[2] Bonnel, S., Oury C. 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. Dans les *Actes de la 80e conférence de l'IFLA* (Lyon, France, août 2014). [http://library.ifla.org/998/1/107-bonnel-fr.pdf].

[3] Oury C. 2011. When *press* is *not printed:* the challenge of collecting digital newspapers at the Bibliothèque nationale de France. Dans les *Actes de préconférence de la section presse de l'IFLA* (Mikkeli, Finlande, août 2012). [http://www.ifla2012mikkeli.com/getfile.php?file=154 or http://halshs.archives-ouvertes.fr/docs/00/76/90/84/PDF/LegalDepositNewspapersBnF_Oury_IFLA2012.pdf]