Bibliothèque nationale de France

Accès à distance aux archives de l'internet dans les BDLI : mémento

Bibliothèque nationale de France

direction des Services et des réseaux
département du Dépôt légal
dépôt légal numérique

TABLE DES MATIERES

1. INTRODUCTION ET OBJECTIFS 3
2. CONTEXTE ET ASPECTS REGLEMENTAIRES 3
2.1. LA MISSION DE DEPOT LEGAL DE L'INTERNET 3
2.2. LES MODALITES DE CONSULTATION DES COLLECTIONS 3
2.3. LE ROLE DES BIBLIOTHEQUES DE DEPOT LEGAL IMPRIMEUR 3
3. ASPECTS TECHNIQUES 4
4. ASPECTS ORGANISATIONNELS 4
5. COLLECTIONS 4
6. UTILISATION DES COLLECTIONS PAR LES LECTEURS 5
7. OUTILS DE RECHERCHE 6
8. CONTACTS A LA BNF 8
9. EN SAVOIR PLUS 8

1. Introduction et objectifs

Ce mémento fournit des éléments d'accompagnement à la mise en place d'un accès distant aux archives de l'internet de la BnF dans les Bibliothèques de dépôt légal imprimeur (BDLI). Il est destiné aux agents susceptibles de faire de la médiation auprès des lecteurs, aux encadrants et aux équipes impliquées dans le dispositif. Le document cherche à résumer les aspects importants, techniques, organisationnels ou documentaires, pour permettre aux équipes de répondre aux questions des utilisateurs et intégrer le service d'accès à distance dans l'offre de la bibliothèque.

Ce document ne concerne que l'accès aux collections du dépôt légal de l'internet ; la sélection des sites par les BDLI, ainsi que le dépôt légal imprimeur, sont traités dans des documents à part.

2. Contexte et aspects réglementaires

2.1. La mission de dépôt légal de l'internet

La BnF a pour mission de collecter, conserver et communiquer les sites internet du domaine français au titre du dépôt légal (Code du patrimoine - Titre III : Dépôt légal - Partie législative (Article L131-1 – L132-4) et Partie réglementaire (R132-23)).

L'internet « français », soumis au dépôt légal de la BnF, est constitué par :

- l'ensemble des sites en .fr ou autres extensions liées au territoire (.re., .nc., etc.) ;
- les sites dont les producteurs sont domiciliés en France ou dont les contenus sont produits en France, quel que soit leur nom de domaine.

La collecte des sites internet des chaînes de télévision et de radio françaises ainsi que des sites qui y sont « principalement consacrés » est cependant assurée par l'Institut national de l'audiovisuel (INA).

2.2. Les modalités de consultation des collections

Les archives de l'internet constituent un fonds patrimonial qui relève du dépôt légal. Les règles appliquées pour l'accès aux archives à la BnF sont identiques à celles des autres collections de dépôt légal, et leur accès est limité aux chercheurs accrédités.

L'accès à distance est autorisé selon l'article R132-23-2 du Code du patrimoine, qui précise :

> *La consultation sur place des services de communication au public en ligne et des services de médias audiovisuels à la demande collectés s'effectue :*
>
> *1° A la Bibliothèque nationale de France et dans tout organisme habilité à mettre en œuvre cette consultation par arrêté du ministre chargé de la culture ;*
>
> *2° Sur des postes individuels équipés d'interfaces d'accès, de recherche et de traitement fournies par la Bibliothèque nationale de France ou les organismes habilités et dont l'usage est strictement réservé à des chercheurs dûment accrédités.*

L'arrêté ministériel du 16 septembre 2014 établit la liste des 26 bibliothèques du dépôt légal imprimeur habilitées à mettre en place un accès à distance aux collections du dépôt légal de l'internet. Le dispositif technique mis en place par la BnF permet d'établir cet accès ; il est encadré par un conventionnement avec la BnF. Les aspects techniques et organisationnels de cet accès sont décrits en section 3.

2.3. Le rôle des Bibliothèques de dépôt légal imprimeur

Les BDLI peuvent être impliquées aussi bien dans la valorisation des archives de l'internet que dans la sélection de sites à collecter. Plusieurs BDLI ont déjà contribué à la sélection de sites lors des collectes électorales menées depuis 2004. La BnF travaille avec les BDLI en réseau et à la mise en place de collectes pilotées par les BDLI de sites en lien avec leur région. La politique documentaire de chaque collecte est définie par la BDLI en lien avec la BnF ; la première collecte, celle des « Alsatiques en ligne », a été créée en 2013 à l'initiative de la BNU de Strasbourg.

En ce qui concerne la valorisation, les BDLI sont responsables de la promotion des archives du web et de l'accompagnement de leurs utilisateurs. La BnF travaille avec les équipes sur place pour fournir l'information et le

support technique nécessaires. Les BDLI peuvent également choisir d'élaborer des « parcours guidés » sur leur sélection ; la mise en ligne de ces parcours sur l'application de consultation des archives est programmée avec la BnF. La section 4 du présent document fournit des éléments sur le contenu et le fonctionnement des archives de l'internet.

3. Aspects techniques

Le dispositif informatique mis en place par la BnF permet d'intégrer la consultation des archives de l'internet sur un poste informatique appartenant à la BDLI ; ce poste n'a pas besoin d'être dédié à cette seule consultation. Le dispositif nécessite l'installation du plug-in *Digital DNA* édité par la société LoginPeople. Ce plug-in permet d'enregistrer préalablement puis d'authentifier les postes se connectant au système d'information de la BnF. Les lecteurs peuvent ensuite utiliser un navigateur, aller à l'adresse https://dli.bnf.fr et avoir accès aux archives de l'internet après avoir accepté la charte du bon usage des services à distance de la BnF. Cette session de navigation s'effectue dans un environnement sécurisé, sur un poste virtuel hébergé sur les serveurs de la BnF, à l'aide d'un deuxième navigateur identifié par une couleur orange et qui se trouve imbriqué dans le navigateur local.

Pré-requis : Le dispositif est compatible avec les systèmes Windows et Mac OS. Il nécessite un accès internet, ainsi qu'un navigateur récent compatible avec les technologies HTML5 et WebSockets. L'installation du plug-in et l'enregistrement du poste nécessitent des droits d'administration.

4. Aspects organisationnels

Chaque BDLI bénéficiant du dispositif dispose de deux postes d'accès à distance : un poste public et un poste professionnel.

Conformément au Code du patrimoine, l'accès aux archives de l'internet est limité aux chercheurs accrédités. Le processus d'accréditation des chercheurs peut être intégré à un processus existant pour d'autres fonds patrimoniaux dans la BDLI. Pour les demandes d'accès spécifiquement aux archives de l'internet, la BnF applique le principe suivant :

> *Toute personne justifiant du besoin de consulter les archives pour des raisons d'études, universitaires, professionnelles ou personnelles, peut obtenir un titre d'accès correspondant au temps nécessaire à sa recherche.*

Le poste public qui donne accès aux archives de l'internet peut par exemple se situer dans un espace dont l'accès est contrôlé ; la connexion au poste peut aussi être protégée par un mot de passe attribué par les agents de la BDLI. Les détails sont définis en collaboration entre la BnF et la BDLI en fonction de la situation dans chaque bibliothèque.

5. Collections

Pour archiver les sites internet, la BnF réalise des collectes automatiques à l'aide de robots moissonneurs qui copient pages, images, animations, fichiers audio et vidéo. Les fichiers des sites sont ensuite datés et indexés pour être restitués dans leur contexte de publication original, ce qui permet de naviguer dans les archives comme sur l'internet, en cliquant de lien en lien.

Compte tenu de la masse d'informations disponibles sur l'internet et des techniques de publication utilisées par les éditeurs de sites, tous les sites et toutes les pages des sites ne peuvent être archivés. La BnF constitue des échantillons représentatifs de l'internet français en combinant deux modes d'entrée :

- les **collectes larges** permettent de constituer des échantillons représentatifs du web (4,1 millions de sites en 2013). Réalisées une fois par an, elles portent aujourd'hui principalement sur les domaines .fr et .re, grâce à un partenariat avec l'Association française pour le nommage internet en coopération (AFNIC) et sur les domaines .nc, grâce à un partenariat avec l'Office des postes et télécommunications de Nouvelle-Calédonie (OPT-NC).
- les **collectes ciblées** portent sur une sélection d'environ 20 000 sites repérés par des bibliothécaires. Ces sites sont choisis en raison de leur thème (la littérature, le développement durable…) ou de leur rapport à un événement (comme les élections ou les Jeux Olympiques en 2012). Les collectes ciblées sont soit plus profondes (pour archiver les grandes bases documentaires), soit plus fréquentes (une centaine de journaux en ligne font ainsi l'objet d'une collecte quotidienne afin de saisir l'actualité du web).

Ces collectes sont complétées par des collections « historiques » acquises auprès de la fondation américaine Internet Archive ; les plus anciennes captures dans les archives datent de 1996. Les captures plus récentes sont en général plus complètes en raison de l'évolution du dispositif de collecte à la BnF.

En consultant les archives on peut constater que certains contenus (pages, images, vidéos…) sont absents. Les archives de la BnF sont lacunaires pour plusieurs raisons :

- en raison du modèle de collecte : la BnF ne peut pas collecter de façon exhaustive l'ensemble des documents mis en ligne, mais elle cherche à constituer des échantillons représentatifs de l'internet français en associant des collectes larges et des collectes ciblées,
- parce que le robot moissonneur utilisé par la BnF pour constituer les archives a des limites : il peut rencontrer divers obstacles en parcourant l'internet (indisponibilité des serveurs, pièges, redirections...) et ne collecte pas systématiquement les contenus payants ou sur mot de passe, les flux audio et vidéo, les animations interactives ou contenus accessibles par formulaires (moteurs de recherche, annuaires, catalogues en ligne),
- parce que les outils de consultation ont aussi des limites : ils ne sont pas capables de restituer entièrement et correctement toutes les fonctionnalités mises à disposition des internautes.

Il n'existe pas de liste complète des sites archivés par la BnF. Cependant, les fiches thématiques de data.bnf.fr signalent les sites web sélectionnés par la BnF dans le cadre des collectes ciblées. Par exemple : http://data.bnf.fr/11932277/litterature_francaise/. Certains sites font aussi l'objet de notices dans le catalogue général de la BnF ; pour l'instant il s'agit uniquement de titres de presse collectés quotidiennement par la BnF.

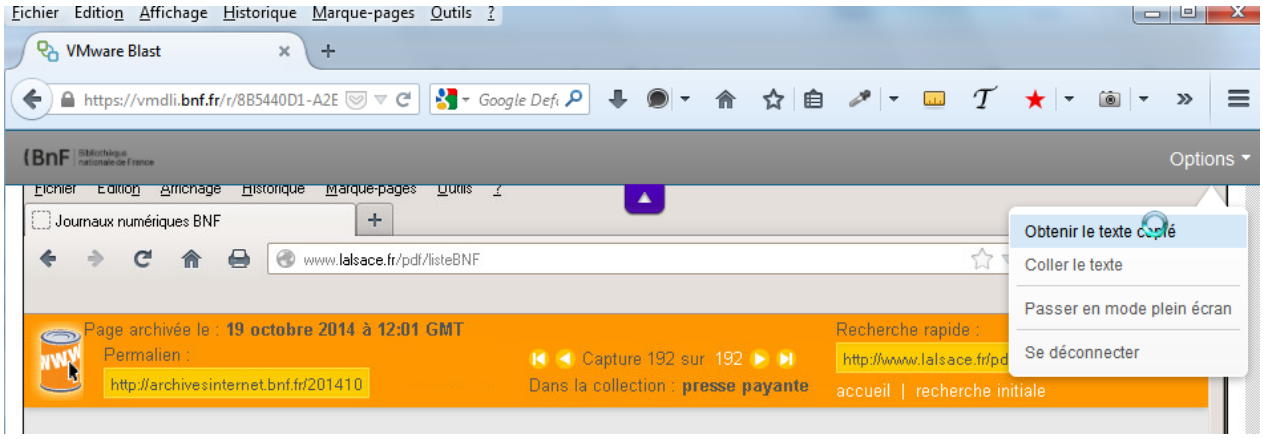
6. Utilisation des collections par les lecteurs

L'accès aux collections se fait 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 navigateur Firefox de couleur orange 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 un logiciel libre appelé Wayback.

Dans le respect de la loi, il est interdit de télécharger les fichiers archivés (textes, images, vidéos…). Le dispositif informatique empêche le téléchargement sur le poste local.

En revanche, sous réserve des règles et des outils mis en place dans les BDLI, il est possible de copier des extraits de texte ou faire des copies d'écran à des fins de citation.

Fonctions Copier / coller



Les fonctions de « copier/coller » sont possibles mais sont rendues plus complexes par l'utilisation d'un poste virtuel. Pour les utiliser, il faut passer par le menu « Options » de la session de navigation qui est accessible en cliquant sur la flèche située en haut et au centre de la fenêtre du navigateur.

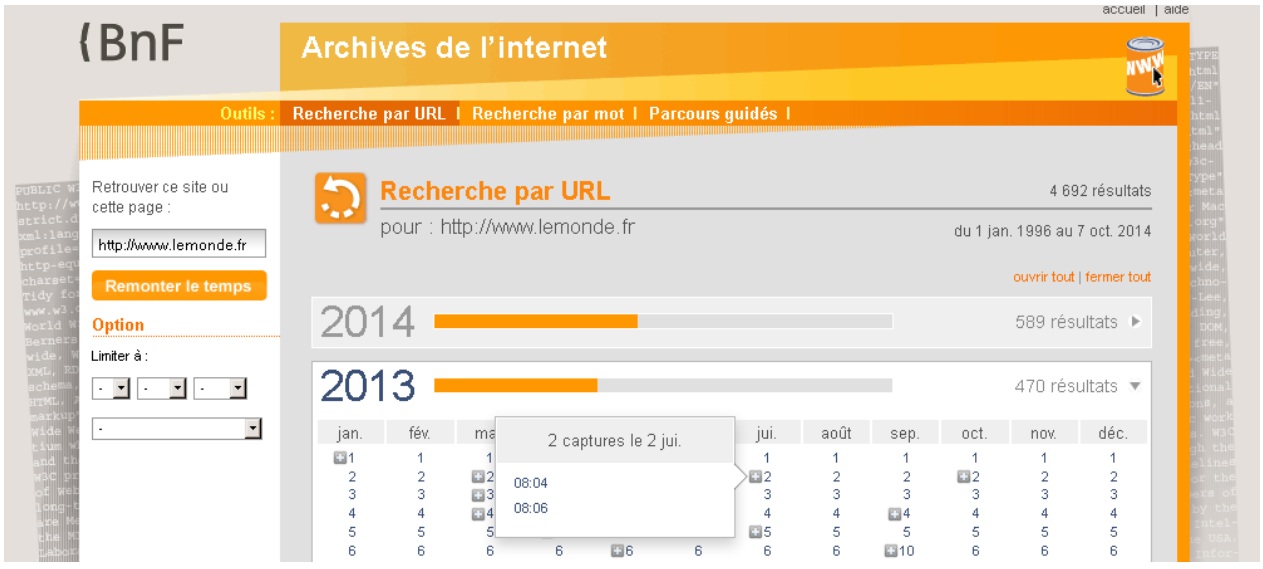
- Pour **récupérer un extrait de texte archivé**, sélectionner le texte avec la souris, utiliser les menus (menu Edition ou menu contextuel sur le clic droit) du navigateur orange pour mettre le texte dans le presse-papier du poste virtuel. Choisir ensuite l'option « Obtenir le texte copié » et suivre les explications.
- Pour retrouver un site archivé en cherchant son adresse URL, il est possible de **coller l'URL dans un champ de recherche.** Copier l'URL dans le presse-papier local (en utilisant les menus ou les raccourcis). Choisir ensuite l'option « Coller le texte » et suivre les explications.

7. Outils de recherche

Les archives sont accessibles selon deux modes principaux de recherche : la recherche par URL et les parcours guidés.

La **recherche par URL** : elle permet de retrouver l'archive d'un site, d'une page ou d'un fichier en indiquant son adresse internet (son URL, de l'anglais *Uniform Resource Locator*).

Par exemple, saisir l'adresse http://www.lemonde.fr dans le formulaire de recherche permet de retrouver toutes les dates auxquelles le site du journal *Le Monde* a été archivé. Il suffit ensuite de choisir une date et de cliquer de lien en lien pour consulter l'archive constituée à cette date. Le signe « + », qui apparait à côté de la date, indique qu'il y a plusieurs captures dans la journée.



Attention : toutes les pages d'un même site ne sont pas systématiquement archivées ; elles ne sont pas non plus archivées de façon simultanée : en cliquant sur un lien, l'application tente d'afficher la page dont la date d'archivage est la plus proche de son point de départ (celle de la page consultée).

La recherche par URL porte sur la totalité des archives constituées par la BnF (21,2 milliards de fichiers, 468 To de données au 31/12/2013).

Des options de recherche avancée permettent de limiter la recherche à une date ou une période d'archivage, ou de limiter la consultation à une collection particulière.

On conseille aux lecteurs de naviguer entre l'internet « vivant » (navigateur bleu) et les archives de l'internet (navigateur orange) pour retrouver l'adresse URL d'un site internet et voir son évolution au fil des années.

Les **parcours guidés** : destinés à mieux faire connaître les collections et à familiariser les lecteurs avec les outils de navigation, ils permettent d'explorer les archives sur des thèmes choisis par les agents de la BnF ou des chercheurs partenaires. Il y a aujourd'hui dix parcours guidés ; de nouveaux parcours sont rajoutés régulièrement :

- Presse et actualité
- Commémorer en ligne : Jean-Philippe Rameau (1683-1764)
- Le web scientifique : de la vulgarisation aux sciences participatives
- Carnets de voyage : le monde au bout des doigts
- L'administration en ligne : le web au service des citoyens
- Images amateurs, amateurs d'images
- La révolution tunisienne à travers le web
- Le web vert : les politiques du développement durable
- Le web militant
- (S') écrire en ligne : journaux personnels et littéraires
- Cliquer, voter : l'internet électoral

Il existe aussi une **recherche par mot** mais elle est très expérimentale et ne porte que sur une petite partie des collections, les collectes larges de 2006 et 2007. Elle peut servir comme point d'entrée dans les archives mais son utilité est limitée. La BnF souhaite travailler sur l'amélioration de ce service dans les années à venir.

Lors de la consultation des archives le **bandeau orange** en haut de la page affiche la date et heure exactes d'archivage de la page. A noter que les archives ne permettent pas de connaître la date de publication ou de mise en ligne d'une page web : même si certains éditeurs de sites choisissent de l'afficher, il est impossible de garantir son authenticité. Mais les archives permettent de savoir qu'une page était accessible aux internautes à la date à laquelle le robot d'archivage est passé.

Enfin, un système de **permalien** permet de citer précisément une URL archivée à une date et une heure données. Pour accéder à cette capture ultérieurement, il suffit de copier-coller le permalien dans la barre d'adresse du navigateur orange.

Problèmes et messages d'erreur

Les usagers peuvent rencontrer des difficultés ou voir des messages d'erreur en consultant les archives de l'internet, voici la liste des cas les plus fréquents :

- Message « Cette page n'est pas dans les archives » : indique que la page n'a pas été collectée, soit parce qu'elle ne faisait pas partie du périmètre de la collecte, soit parce que le robot a rencontré des difficultés qui ont empêché la collecte.
- Message « Cette page ne peut momentanément pas être affichée pour des raisons techniques » : indique que la page a été collectée mais qu'il y a un problème d'accès aux données. Si le message s'affiche de manière persistante, il est utile de le signaler à la BnF.
- Problèmes de mise en page, zones vides sur la page, écran noir/blanc : ces phénomènes indiquent qu'il y a eu un problème soit lors de la collecte, soit lors de l'affichage. Ces problèmes concernent souvent les vidéos, les éléments en Flash ou javascript, les pages nécessitant une recherche en base de données, les feuilles de style.
- Pages avec des codes HTTP de type 404, 503, etc. : en général ces messages ont été collectés par le robot car ils représentent l'état du site au moment où le robot est passé. Dans certains cas ils peuvent être produits par le comportement du robot (un site qui a bloqué l'accès, des URL fausses générées par le robot) mais d'autres résultent des liens cassés, contenus manquants ou problèmes serveur sur le site.

8. Contacts à la BnF

Pour des questions sur le dispositif technique ou des problèmes de connexion : Sara Aubry, département des Systèmes d'information, sara.aubry@bnf.fr

Pour des questions sur les collections ou des questions générales sur l'accès à distance : Ange Aniesa, département du Dépôt légal, ange.aniesa@bnf.fr

9. En savoir plus

Sur les collections du dépôt légal de l'internet :
http://www.bnf.fr/fr/collections_et_services/livre_presse_medias/a.archives_internet.html

Sur les modalités du dépôt légal des sites web :
http://www.bnf.fr/fr/professionnels/depot_legal/a.dl_sites_web_mod.html

Sur les processus et techniques mis en place par la BnF :
http://www.bnf.fr/fr/professionnels/innov_num_dl_internet.html