|  |  |  |  |
| --- | --- | --- | --- |
|  | | Bibliothèque nationale de France  direction des Services et des réseaux  département du Dépôt légal  dépôt légal numérique | |
| 11/06/2018 15:06:15 | | Quai François-Mauriac 75706 Paris Cedex 13 | |
|  | | téléphone | 33 (0)1 53 79 59 28 |
|  | | télécopie | 33 (0)1 53 79 85 56 |
|  | |  |  |
| Réf : BnF-ADM-2018-061898-01 | Affaire suivie par : Marie CHOULEUR  tél : 33 (0)1 53 79 54 93 | **note** | |
| **note de** | **Marie CHOULEUR** |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
| **à l’attention de** | **Benoît TULEU** |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
| **date** | **11 juin 2018** |  | |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **objet** | **Pour une définition du domaine français de l'internet et le périmètre de collecte au titre du dépôt légal : synthèse de la pré-étude confiée à Camille Tatger** | |
| Votre dossier |  | |

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | |
|  | |  |  | | | |  |
| Copies | | DDL\_DLN, DDL\_CSA, DDL\_CM | | | | | | |

Le service du dépôt légal numérique a accueilli du 2 au 25 mai 2018 Camille Tatger, élève conservatrice du patrimoine dans le cadre de son stage hors spécialité à l’Institut national du patrimoine.

L’objectif de la pré-étude qui lui a été confiée était d’initier une réflexion sur la mise en œuvre de sa mission de dépôt légal de l’internet français par la BnF et de formuler des propositions.

Ces travaux ont permis de valider des intuitions en leur donnant une assise chiffrée, et de dégager de premières orientations pour la conduite de l’édition 2018 de la collecte large.

Un accompagnement par la délégation à la Stratégie ou un prestataire extérieur permettrait au service de prolonger la réflexion fondamentale engagée à l’occasion de ce stage.

Non contente de ré-interroger les pratiques de la BnF, cette pré-étude a mis en évidence la nécessité d’expliciter les choix, scientifiques ou techniques, ayant présidé à la constitution des collections afin qu’ils soient intelligibles par les chercheurs de demain.

***Ce que disent les textes***

Le dépôt légal de l’internet français est encadré par la loi DADVSI de 2006 et de son décret d’application de 2011, codifiés dans le Code du patrimoine.

Il vise les **contenus publiés** en dehors du cercle de la famille par **voie électronique** ressortissant du **domaine français**, entendu comme les services de communication au public, ou sites internet :

* enregistrés sous le nom de domaine. fr ou tout autre nom de domaine enregistré auprès du ou des organismes français chargés de la gestion de ces noms ;
* enregistrés par une personne domiciliée en France ;
* ou produits sur le territoire français.

La BnF est tenue de réaliser, *a minima*, une **collecte annuelle** de ces contenus. Elle peut solliciter auprès des éditeurs la **fourniture de mots de passe et clés d’accès** aux documents protégés.

***Ce que fait la Bibliothèque nationale de France***

La BnF fait intervenir un principe de **représentativité** qui n’apparaît pas dans les textes. Son modèle de collecte, similaire à celui des autres bibliothèques nationales qui archivent le web à travers le monde, repose sur la **complémentarité** entre :

* une collecte « large », annuelle, à partir de listes de sites recueillies par l’intermédiaire de bureaux d’enregistrement partenaires, avec un budget, c’est-à-dire un nombre de fichiers collectés, identique pour tous les sites ;
* et des collectes « ciblées », à partir de listes de sites constituées par des bibliothécaires, et organisées de manière thématique, avec des paramètres de fréquence, de profondeur et de budget variables.

Le seuil d’**accroissement annuel** a été fixé à **150 To**, à répartir entre la **collecte large**, qui porte sur près de **4,5 millions de sites**, et les **collectes ciblées** qui concernent **quelques dizaines de milliers de sites**.

La BnF sollicite les éditeurs uniquement dans le but de collecter certains **titres de presse** ou **publications** car les échanges initiaux, les instructions préalables et les opérations de maintenance sont coûteuses, et sans garantie de succès.

Elle s’interdit de collecter les **réseaux sociaux**, à l’exception des comptes des personnalités publiques et des organisations, et de rares hashtags, les **intranets** et les **contenus en accès protégé**, sauf cas mentionnés ci-dessus.

Proposition n°1: Etudier l’opportunité de solliciter un nouvel avis de la CNIL pour préciser le statut des contenus échangés publiquement sur les réseaux sociaux.

***Faire progresser la couverture du domaine français***

Au-delà de ces **restrictions de principe**, la collecte par la BnF se heurte à des **limites**.

1. Concernant la collecte large

Le recours aux bureaux d’enregistrement est le seul moyen qui permette à la BnF d’identifier, à un niveau macro, les sites internet ressortissant du domaine français. Pour autant, rien ne contraint ces acteurs à collaborer.

La gestion des noms de domaine est complexe, avec des risques de recouvrement : en France, 18 opérateurs sont accrédités par l’ICANN. Il faut donc conduire un travail d’identification et de persuasion de ces acteurs.

Au terme de cette pré-étude, Camille Tatger a évalué le périmètre du domaine français de l’internet à 8 millions de noms de domaine :

* 3,5 millions d’entre eux sont enregistré en .fr, et sont collectés par la BnF en vertu de son partenariat avec l’Afnic ;
* 400 000 le sont sous de nouvelles extensions, et sont pour partie collectés par la BnF, qui mène une veille sur le sujet ;
* et environ 4 millions le sont sous les extensions génériques, telles que .com ou .net.

Si ces dernières représentent près de la moitié des noms de domaine enregistrés, elles constituaient en revanche moins du tiers des domaines moissonnés lors de l’édition 2017 de la collecte large. La société Gandi enregistre à elle seule la moitié de ces noms de domaine.

Proposition n°2 : Augmenter la proportion de noms de domaine enregistrés sous des extensions génériques dans la collecte large par des partenariats.

1. Concernant les collectes ciblées

La sélection procède d’une logique acquisitive, puisque les sites archivés ont vocation à s’inscrire dans la continuité thématique des collections imprimées. Elle se trouve limitée par des facteurs organisationnels et humains.

Elle est orientée par la politique documentaire des départements, et par la sensibilité des sélectionneurs, et non par les usages de l’internet : ainsi, la notoriété et le trafic des sites ne constituent pas des critères de sélection.

Par ailleurs, les alternatives françaises aux sites étrangers ne sont pas systématiquement mises en exergue : AdopteUnMec, Dailymotion, Deezer, KissKissBankBank, etc.

La pré-étude a montré que leurs principaux usages concernaient la communication (notamment au travers des réseaux sociaux), la vente en ligne et l’e-administration, devant la consultation des sites de presse et d’information.

Proposition n°3: Promouvoir de nouvelles méthodes de veille, associant approche quantitative et approche qualitative, auprès du réseau des correspondants.

Proposition n°4: Développer la sélection partagée en coopération avec les bibliothèques de dépôt légal imprimeur en région et les acteurs du monde de la recherche.

1. Pour tous types de collecte confondus

La BnF a choisi d’internaliser l’activité et d’adopter une approche mono-robot, basée sur l’utilisation de logiciels *Open Source* et de formats normalisés, pour garantir la maintenabilité de la chaîne, et la pérennité des données produites.

Dans ces conditions, le succès de la collecte est dépendant de l’architecture des sites et des technologies utilisées : bases de données, cartes interactives ou « long scrolling pages », par exemple, sont parmi les contenus qui échappent au robot.

Les captchas et autres dispositifs de blocage des robots peuvent faire obstacle à la collecte. A la marge, il arrive que les éditeurs adressent à la BnF des réclamations ou ne lui permettent pas d’accéder aux contenus en accès protégé

On constate que les usages en mobilité se développent : 60 % des Français ont accès à l’internet sur leur téléphone portable, et 85% grâce à leur ligne fixe. Or, la collecte porte sur les sites classiques et non sur les versions mobile ou les applications.

Proposition n°5 : Documenter grâce à des didacticiels, des sessions de navigation filmées ou par tout autre moyen le parcours des internautes et la dynamique des sites.

|  |  |
| --- | --- |
|  | Marie CHOULEUR |
|  | chef de service |