BnF | Bibliothèque nationale de France

Archivage du web français par la BnF : 6 milliards d'URL collectées en 2023

Contacts

Élodie Vincent
Cheffe du service de presse et des partenariats médias
01 53 79 41 18

Pierre Clamaron
Chargé de communication presse et partenariats médias
06 59 08 81 57

presse@bnf.fr

Les collections du dépôt légal du web, que la Bibliothèque nationale de France constitue depuis 2002 dans le cadre de sa mission de conservation patrimoniale, ont franchi en 2023 le seuil des 2 pétaoctets (soit 2 000 téraoctets) de données.
La collecte large annuelle et les collectes ciblées réalisées par la BnF ont permis de sauvegarder 4,4 milliards d'URL, auxquelles s'ajoutent deux collectes « d'urgence » de plateformes ayant fermé à l'été 2023 : les skyblogs et les pages personnelles Orange, pour un volume de 1,9 milliard d'URL.
C'est une matière exceptionnelle, tant par son contenu que par son volume, qui est ainsi mise à disposition des chercheurs par la Bibliothèque.

S'inscrivant dans la continuité du dépôt légal des documents déjà collectés (livres, journaux, revues, disques, vidéos et jeux vidéos…), le dépôt légal de l'internet a été amorcé par la BnF en 2002, et archive des sites mis en ligne à partir de 1996. Il s'applique à toutes les publications du web français.

À l'aide de logiciels d'archivage automatique en ligne, **la BnF réalise des « moissonnages » de l'internet français une fois par an lors de sa collecte annuelle, mais aussi plus régulièrement à l'occasion de collectes ciblées, en lien avec les collections thématiques et spécialisées de ses départements ou avec l'actualité** nationale et internationale (guerre en Ukraine, échéances électorales, Jeux olympiques…).

Compte tenu des masses de données en jeu, cet archivage ne prétend pas à l'exhaustivité, mais vise à assurer la **meilleure représentativité possible du web français**.

Le dépôt légal de l'internet garantit le respect du droit de la propriété intellectuelle, en donnant accès aux collections archivées à des fins de recherche et à des lecteurs accrédités, exclusivement dans les emprises de la BnF et celles de ses partenaires en région et en outre-mer.

Son exercice est réalisé en **conformité avec les recommandations de la CNIL** (Commission nationale de l'informatique et des libertés) **relatives à la protection des données personnelles**.

La **collecte annuelle 2023** s'est déroulée du 18 octobre au 5 décembre et a porté sur 5 731 808 domaines de départ à raison de 2 200 URL collectées par domaine. **3 173 362 231 URL ont ainsi pu être sauvegardées.**

La BnF a par ailleurs poursuivi l'**intégration de nouveaux contenus tels que les réseaux sociaux** (YouTube, Instagram, TikTok) ou les **podcasts**, même si le passage de Twitter à X et les nouvelles modalités d'accès associées à ce réseau social ne rendent plus possible sa collecte par les robots de la BnF depuis le mois de juin 2023.

Deux collectes d'urgence se sont, en outre, ajoutées aux activités courantes du dépôt légal du web en 2023, pour préserver les contenus de deux plateformes majeures qui avaient annoncé leur fermeture en milieu d'année :

La collecte des **Skyblogs**, l'un des premiers réseaux sociaux lancé en 2002 qui mettait gratuitement à disposition de ses membres un espace numérique personnalisé.
Cette collecte a duré **85 jours**, du 28 août au 17 novembre, et a permis de sauvegarder **12 607 289 blogs** pour un total d'**URL** collectées s'élevant à **1 873 993 846** (dont 1 093 089 908 images et 729 475 996 pages web).

La collecte des **pages personnelles Orange**, espace qui permettait aux clients d'Orange de créer un ou plusieurs sites internet de manière assistée ou autonome.
Cette dernière a eu lieu du 20 novembre au 7 décembre et a permis la sauvegarde de **298 188 sites et 26 094 982 URL.**

L'intégralité des contenus archivés grâce à ces collectes est désormais accessible aux chercheurs, à la BnF ainsi que dans ses bibliothèques partenaires en région et en outre-mer, et ce, dans le respect du droit de la propriété intellectuelle.

En 2024, de nouvelles collectes ciblées, portant sur les Jeux olympiques ou encore les élections européennes, viendront encore enrichir les milliards d'URL déjà conservées par la Bibliothèque.