Une archive pour la recherche :
Web et coronavirus (COVID-19)

La BnF a la charge du dépôt légal de l'internet français. Sa collection de sites archivés, parmi les plus anciennes et les plus riches au monde, constitue une source inédite pour les chercheurs en sciences humaines et sociales. Elle fait l'objet depuis quelques années d'importants projets de recherche qui mobilisent les techniques issues de l'intelligence artificielle, dont la fouille de données (*text, data and link mining*) : cartographie du web de la Grande Guerre (projet « Le devenir en ligne du patrimoine numérisé »), analyse de la médiatisation des mémoires de l'immigration maghrébine, du développement des néologismes (projet « Néonaute »), etc.

En prise directe avec l'actualité, les équipes de la BnF chargées du dépôt légal numérique rassemblent la matière sur laquelle travailleront les chercheurs de demain.

De janvier à juillet, elles ont procédé à une collecte d'urgence liée à la crise sanitaire du Covid-19. Cette collecte, décrite ci-dessous, constitue une archive unique pour comprendre la crise et ses discours.

La BnF est ouverte à tout partenariat ou projet de recherche sur cette archive.

Contenu

Type de collecte

Collecte d'urgence à partir d'une sélection d'URL effectuée par des bibliothécaires de la BnF et des correspondants dans 15 établissements partenaires en région (BDLI).

Mots-clés

Crise ; Coronavirus (covid-19) ; Épidémie ; Médecine ; Société ; Économie ; Politique ; Confinement

Période

15 février 2020 / 15 juillet 2020 (date de sélection du contenu), soit des premiers cas déclarés en France jusqu'à la fin de l'état d'urgence sanitaire.

Langue(s)

Français, ponctuellement anglais

Volumétrie

Environ 12 téraoctets (dont 1 To de vidéos), en données compressées, à partir d'une sélection de plus de 5000 URL.

La liste des URL de départ et les paramètres de collecte sont communicables et seront versés sur le site api.bnf.fr.

Si le fonds est clos au 15 juillet 2020, les archives sur le sujet continuent de s'accroitre via l'actualité du web et sa collecte courante, notamment concernant les territoires d'outre-mer et la période de déconfinement.

Description du contenu

Les pages préservées rendent compte du caractère global de la crise, de sa couverture médiatique, des prises de position des différents acteurs institutionnels mais aussi de simples citoyens, et plus largement des actions mises en œuvre à travers le web pour endiguer la pandémie, la documenter et la comprendre.

Le cœur de la collection comprend deux ensembles documentaires spécifiques à l'évènement. Le premier correspond aux sélections menées par les bibliothécaires de la BnF et le réseau des correspondants régionaux. Ces sélections couvrent l'ensemble du territoire national et une grande variété de médias : sites des grandes institutions, de presse quotidienne, de presse professionnelle, réseaux sociaux et sites personnels. Pour une même institution, les captures à intervalle régulier des pages dédiées au covid-19 permettent de conserver plusieurs états et rendre compte de l'évolution du site au cours du temps.

Le second ensemble documentaire a été produit à partir de la liste des noms de domaine enregistrés auprès de l'AFNIC et a permis d'archiver une part importante des nombreux sites créées spécifiquement à l'occasion de la crise sanitaire (plate-forme d'entraide, site proposant un diagnostic en ligne, site marchand...).

En dehors de ces deux ensembles spécifiques, on pourra se reporter également aux archives web des élections municipales 2020, aux collections de presse entrées par dépôt numérique et aux collectes courantes menées régulièrement et dont l'objectif est de donner une vision représentative du web français et dont les archives permettent de suivre sites web et comptes de réseaux sociaux dans la durée.

Une collecte vidéo axée sur des chaines présentant un fort lien avec le covid-19 est en cours de préparation.

Accès et reproduction

Consultation sur place (espace recherche de la BnF) et dans les établissements partenaires de la BnF sur poste dédié.

Pour des besoins de recherche, possibilité de fournir les fichiers de conservation au format WARC[1] sur une infrastructure de stockage sécurisée au sein du DataLab de la BnF.

La collecte dans les médias

Le Monde, Comment les archivistes de la BNF sauvegardent la mémoire du confinement sur Internet :

https://www.lemonde.fr/pixels/article/2020/05/15/a-la-bnf-les-archivistes-du-web-sauvegardent-l-internet-francais-du-confinement_6039704_4408996.html

RFI, La BnF conserve la mémoire du coronavirus sur le Web :

http://www.rfi.fr/fr/culture/20200519-publi-mercredi-matinla-bnf-conserve-la-m%C

[1] The WARC (Web ARChive) format specifies a method for combining multiple digital resources into an aggregate archival file together with related information. The WARC format is a revision of the Internet Archive's ARC File Format [ARC_IA] format that has traditionally been used to store "web crawls" as sequences of content blocks harvested from the World Wide Web. The WARC format generalizes the older format to better support the harvesting, access, and exchange needs of archiving organizations. Besides the primary content currently recorded, the revision accommodates related secondary content, such as assigned metadata, abbreviated duplicate detection events, and later-date transformations.
A WARC format file is the concatenation of one or more WARC records. A WARC record consists of a record header followed by a record content block and two newlines; the header has mandatory named fields that document the date, type, and length of the record and support the convenient retrieval of each harvested resource (file). There are eight types of WARC record: 'warcinfo', 'response', 'resource', 'request', 'metadata', 'revisit', 'conversion', and 'continuation'. The content blocks in a WARC file may contain resources in any format; examples include the binary image or audiovisual files that may be embedded or linked to in HTML pages.

3%A9moire-coronavirus-le-web

Les archives web du Coronavirus, une entreprise collective :

https://webcorpora.hypotheses.org/856

The French coronavirus (COVID-19) web archive collection: focus on collaborative networks:
https://netpreserveblog.wordpress.com/2020/05/27/the-french-coronavirus-covid-19-web-archive-collection/

Documents remarquables ou représentatifs

https://www.pasteur.fr/fr/espace-presse/coronavirus-toute-actualite-institut-pasteur-covid-19

https://dc-covid.site.ined.fr

https://twitter.com/raoult_didier

http://covid-documentation.aphp.fr

http://www.cafepedagogique.net/lesdossiers/Pages/2020/Coronavirus2020.aspx

http://oreilletendue.com/2020/04/15/bref-lexique-du-confinement/

Contacts

Alexandre Chautemps, chef du service du dépôt légal numérique (DDL/DSR) :

alexandre.chautemps@bnf.fr

Peter Stirling, chargé d'appui aux projets scientifiques (DSG) :

peter.stirling@bnf.fr