Fiche de cours

Contenus

Le nombre de sites web à travers le monde augmente de manière exponentielle ; si on dénombrait ainsi près de 1 milliard de sites web en 2016, on en décompte près de 750 millions de plus en 2019. Quelle stratégie d'archivage mettre en œuvre pour un corpus aussi important ? Elle doit reposer à la fois sur une approche et des stratégies de collectes différentes et différenciées.

1. Des approches en matière d'archivage du web différenciées
Quatre approches différentes sont aujourd'hui recensées.

1. Approche intégrale
Il s'agit de collecter la totalité du web, sans distinction, ni critère de sélection. Une seule organisation a mis en œuvre une telle approche : Internet Archive. Elle moissonne ainsi régulièrement la totalité du web (https://archive.org).

2. Approche exhaustive
Il s'agit de viser la complétude dans un périmètre circonscrit ; il peut s'agit d'un domaine spécifique, d'un espace national ou d'une typologie particulière de sites web. La Bibliothèque nationale de France, par exemple, a pour tâche, de par son mandat de dépôt légal du web, de collecter une fois par année la totalité du domaine .fr.

3. Approche sélective
Il s'agit de collecter des sites web en fonction de critères thématiques, qualitatifs…

4. Approche thématique
Il s'agit d'une approche complémentaire de l'approche sélective, qui consiste à sélectionner une collection de sites web dédiés à un événement particulier (un événement sportif, une campagne électorale…).

2. Stratégies de collecte
Sur ces approches se greffent des stratégies de collecte différenciées et différentes.

1. Stratégie automatisée
Il s'agit de mettre en place un logiciel-robot (moissonneur ou collecteur du web). Cette stratégie automatisée est notamment utilisée dans le cadre des approches intégrale et exhaustive.

2. Stratégie semi-automatisée
Cette stratégie recourt non seulement au logiciel-robot, mais fait également appel à des critères de sélection plus précis. Elle est utilisée notamment dans le cadre d'une approche sélective.

3. Approche manuelle
Cette stratégie est mise en œuvre dans le cadre d'une sélection manuelle de sites pertinents. Elle est surtout utilisée dans le cadre d'une approche thématique.

3. Comment s'effectue le moissonnage des sites web

Le moissonnage est effectué par un robot d'indexation. Le plus connu d'entre eux est le robot Heritrix, qui a été conçu et est utilité par Internet Archive. Le robot se positionne sur la page d'accueil du site à moissonner (par exemple, http://www.ebad.ucad.sn/) et il collecte le contenu de toutes les pages du domaine moissonné, ainsi que les liens vers des ressources externes. Internet Archive a également défini un format d'archivage : le format WARC (norme ISO 28500 (https://iipc.github.io/warc-specifications/specifications/warc-format/warc-1.0/)).

De nombreuses institutions présentent de manière exhaustive les modalités de moissonnage qu'elles mettent en œuvre. Parmi celles-ci, figurent :
- la Bibliothèque nationale de France (https://www.bnf.fr/fr/centre-d-aide/depot-legal-des-sites-web-mode-demploi)
- la Bibliothèque nationale suisse (https://www.nb.admin.ch/snl/fr/home/informations-professionnels/e-helvetica/archives-web-suisse.html).

Il convient de relever que l'archivage du web est aujourd'hui réalisé par un consortium de 53 bibliothèques provenant de 45 pays, l'*International Internet Preservation Consortium*. Ce dernier met à la disposition du public de nombreuses ressources :
- une bibliographie (https://netpreserve.org/web-archiving/bibliography)
- des stratégies d'archivage du web (https://netpreserve.org/collection-development-policies)
- des outils et des programmes (https://netpreserve.org/web-archiving/tools-and-software)
- des études de cas (https://netpreserve.org/web-archiving/case-studies)

Même si les bibliothèques nationales réalisent un important travail d'archivage du web, cela ne dispense néanmoins pas les services d'archives de réfléchir à une stratégie de collecte des sites Internet de l'organisation dont ils assurent la conservation à long terme de la mémoire.

Types d'évaluation

- Essai réalisé individuellement
- Présentation d'un dossier en groupe

Ressources bibliographiques

BAUSIRE Jonas, *L'archivage du web : présentation des méthodes de collecte et recommandations pour l'accès aux contenus – et leur structuration –*, dans Revue électronique suisse de science de l'information (RESSI), 2016 (http://www.ressi.ch/num17/article_122)

CHAIMBAULT, Thomas, *L'archivage du web*, Villeurbanne, ENSSIB, 2008 (http://www.enssib.fr/bibliotheque-numerique/documents/1730-l-archivage-du-web.pdf)

CHEBBI Aïda, *Archivage du Web organisationnel dans une perspective archivistique*. Thèse de doctorat, Montréal, EBSI, 2012 (https://papyrus.bib.umontreal.ca/xmlui/handle/1866/9203)