Bibliothèque nationale du Luxembourg
Nationalbibliothéik

L'archivage du web –possibilités et limites

Complétude

L'archivage d'un site web donne la possibilité de capturer l'état du site au moment de l'archivage et de reproduire une copie fidèle à l'original. À l'aide de captures régulières, nous pouvons traverser les différentes versions du même site au fil du temps.

Hélas, il y a de nombreux facteurs qui créent des lacunes et captures incomplètes dans une archive du web. Vu l'évolution rapide des informations en ligne, avec des mises-à jour en continu, il est impossible de capturer tous les changements de tous les sites à tout moment. Les raisons pour ces limites, sont de nature technique et budgétaire. Plus précisément :

- Chaque version archivée d'un site web, montre le contenu qu'on pouvait y retrouver au moment de la capture. Tous les changements qui se déroulent entre deux captures ne feront pas partie de l'archive du web.

- Pour des raisons techniques et budgétaires, nous pouvons seulement effectuer 4 captures par an pour la plupart des sites web en .lu.

- Ces 4 captures sont collectées lors des campagnes à large échelle appelées « domain crawls ». Chaque domain crawl est également limité en termes de la durée totale de 5 semaines et du volume total de données collectées. De plus, les robots-moissonneurs obéissent des lignes directives qui excluent certains contenus :

- La durée limitée de 5 semaines pour l'archivage de tous les domaines, entraîne la possibilité que des sites plus larges et complexes ne seront pas complètement archivés.
- Des fichiers larges, excédant 1 GB, ou nécessitant plus de 20 min pour le téléchargement ne seront pas archivés.
- Des liens menant en dehors d'un domaine en .lu ne seront pas archivés.
- Tout contenu protégé par authentification ne sera pas archivé.
- Tout contenu dynamique (comme Javascript) ou nécessitant une interaction de l'utilisateur (comme un formulaire à remplir, des conditions générales à accepter, etc…) est en tout cas problématique.
- Les fichiers exclus des moteurs de recherche par le standard robots.txt ne sont pas archivés.
- Si le serveur web est configuré avec des limites de téléchargement, l'archivage ne pourra que se faire dans ces limites.
- Si le site web est hors ligne lors du passage du robot, rien ne pourra être archivé.

Certitude

Il est difficile à déterminer si tous les contenus d'un site web ont été capturés. Les limites susmentionnées, ne peuvent pas être dénichées de manière automatique. Il faudrait contrôler manuellement chaque page, chaque document, chaque photo, vidéo ou autre fonctionnalité pour vérifier la complétude de la version archivée. En conséquence, il est risqué de supposer que l'archivage du web est une façon efficace pour archiver tous les contenus d'un site web. Il est possible d'y retrouver les informations qu'on a cherché, mais il est également possible qu'une partie du site soit incomplète ou que certains éléments ne fonctionnent pas correctement.

« Mon site fait partie de l'archive du web,
donc toutes mes photos et documents sont sécurisées »
« Je n'ai plus besoin de garder mon site web en ligne.
Je pourrais toujours le récupérer de l'archive du web. »

Ces suppositions sont problématiques, parce qu'il est impossible de garantir cette certitude, sans examen approfondi, suivi par des retouches et corrections. De manière générale, il est plus fiable de sauvegarder ses documents d'une autre manière que de les stocker sur un site web.

Actions

Avez-vous vérifié, si le site web en question fait
déjà partie de l'archive du web luxembourgeois ?
Vous pouvez trouver un guide pratique sur **webarchive.lu/how-it-works/**

En cas de doute, envoyez-nous les adresses de vos sites web et nous pouvons confirmer s'ils se trouvent sur la liste des moissonnages réguliers : **webharvesting@bnl.etat.lu**

Dans la mesure du possible, nous pouvons offrir notre support pour l'évaluation des causes pour un problème d'archivage ou des captures incomplètes. Il ne nous est cependant pas possible d'offrir des solutions d'archivage spécifiques pour chaque site, dû au grand nombre des sites web dans notre domaine de responsabilité.

Toutes les versions archivées des sites web sont consultables dans l'enceinte de la Bibliothèque nationale.
Pour toute question supplémentaire, veuillez-vous adresser à
Ben Els **Tél.: (+352) 26 559 351**, **webharvesting@bnl.etat.lu**
ou visitez le Luxembourg Web Archive sur **webarchive.lu**