SITES, BLOG ET RÉSEAUX SOCIAUX : QUELQUES SOLUTIONS POUR ARCHIVER LE WEB

Virginien Horge (Archiviste à la Ville de Mons)
version du 14-04-2020

Nous proposons ici d'explorer quelques-unes des solutions d'archivage du web mentionnées dans le manuel pour l'archivage digital à l'usage des particuliers (en néerlandais) et des outils et logiciels proposés par l'International Internet Preservation Consortium (en anglais). Notre présentation reste succincte : n'hésitez pas à tester vous-même ces applications !

Précisions importantes…

Nous sommes loin d'être des informaticiens chevronnés et nous sommes incapables de rentrer dans les détails à ce niveau…

Cependant, un élément est à prendre en compte lorsque l'on décide de télécharger des images, vidéos et textes d'un site web : le droit d'auteur. En effet, librement accessible ne signifie pas libre de droit. Pour une conservation historique et archivistique, l'extraction d'un site internet public n'est pas un problème mais sa réutilisation, par la suite, peut parfois être sujette à discussion. Le mieux est de demander à l'administrateur d'un site internet, d'un blog ou d'une page d'un réseau social s'il peut vous autoriser explicitement à reprendre partiellement son site web pour une conservation à long terme.

Il est important de bien préciser auprès de l'interlocuteur le but de votre démarche : il s'agit de garder son témoignage pour les générations futures mais pas de réutiliser sans vergogne son travail de photographe, par exemple.

Par ailleurs, il est nécessaire de prendre en compte les aspects de vie privée : la présence de commentaire vous empêchera de rediffuser librement de tels enregistrements qui seront cependant utilisables dans le cadre d'une démarche archivistique, historique et scientifique.

Présentations des outils

1. *Enregistrement par un tiers*

Avant tout, la question est de savoir si vous avez les moyens techniques (serveurs, infrastructures, etc.) pour sauvegarder un site web… Ou si vous avez le courage de lancer un tel projet.

N'hésitez donc pas à utiliser le Wayback Machine d'Internet Archives : dans "Save Page Now", vous pouvez suggérer des pages web à enregistrer. Ce site, le plus connu et le plus ancien en termes d'archivage du web, fournit des clichés de pages web, sur plusieurs années, et offre une bonne source pour les sites désormais supprimés ou l'évolution graphique du web.

2. *Un seul poste de réseau social*

Ne cherchez pas à vous compliquer la vie : les solutions les plus simples sont parfois les meilleures… Dans le cadre des réseaux sociaux, peut-être voudrez-vous conserver l'un ou l'autre poste intéressant comme témoignage ou trace d'une activité en ces temps incertains : la fonction "prise d'écran/print screen" vous sera d'une grande aide. Il ne semble pas nécessaire de sortir l'artillerie lourde, pour ces petites traces.

3. *Vous désirez enregistrer une seule page ou un petit nombre de pages d'un site web ?*

a. *Solution sans inscription : deux extensions pour Firefox et Chrome*

Save Page WE
par DW-dev

SingleFile
par gildas

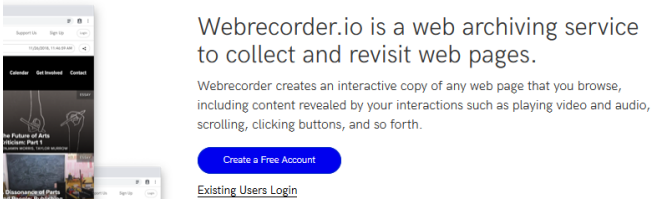
La plupart du temps, une simple extension à ajouter à votre navigateur peut suffire : "SingleFile" ou encore "Save Page WE", deux outils permettant l'enregistrement au format

HTML, en un seul fichier, et existant aussi bien sous Google Chrome/Chromium que sous Firefox.

Par ailleurs, ces extensions fonctionnent également pour l'enregistrement partiel d'une page de réseau social : il faut alors "scroller" jusqu'à atteindre la date désirée (dans notre cas, le début du confinement) puis d'utiliser l'extension. Pour les postes trop longs, il faudra également cliquer sur "afficher la suite".

b. *Solution avec inscription : www.webrecorder.io*

Avec une simple inscription, ce site internet vous permet de facilement créer une collection comprenant l'ensemble des pages que vous désirez.



Les pages, regroupées en collection, s'insèrent dans un même fichier Warc, téléchargeable sur votre ordinateur.

Des fonctionnalités spécifiques existent également pour les réseaux sociaux, YouTube, soundcloud, etc.

c. *Une solution sous Linux : ArchiveBox*

Cette solution permet également d'enregistrer une série de pages sur base d'un dossier de marque-pages ou d'un historique de navigation.

L'interface de lecture est une simple interface en html fournissant l'accès à ces différents formats, avec également un lien vers le site d'origine, la mention d'une date d'enregistrement et de mise à jour éventuelle. Ces métadonnées permettent donc la création d'une interface simple pour l'utilisateur final et les éventuels lecteurs.

L'intérêt de cette solution réside dans la création d'un portail accessible hors ligne comprenant l'ensemble de vos enregistrements et dans la multiplicité des formats d'enregistrement : HTML, PDF, PNG, WARC...

4. Un site complet ou une partie de ce site

a. La solution la plus courante : Httrack

La solution la plus connue et la plus simple à mettre en œuvre est Httrack, open source et disponible gratuitement, utilisable sous Linux mais également sous Windows sans installation. Ce logiciel permet très facilement d'extraire au format HTML et pour une lecture hors ligne un site web, en sélectionnant spécifiquement l'URL du site complet (exemple : www.villeuntel.be/) ou d'une partie de ce site (www.villeuntel.be/santé/covid19). Le logiciel est par ailleurs paramétrable pour la profondeur d'extraction (nombre de sous-menus à extraire, prise en compte des liens vers des sites extérieurs, etc.). Il prend partiellement en compte les aspects dynamiques (vidéos, flash, etc.).

Cette solution fonctionne également pour l'enregistrement d'une seule page.

Il existe un tutoriel en français : https://archive.framalibre.org/IMG/pdf/intro_winhttrack.pdf

b. Une solution sous Linux : Wget

Pour les utilisateurs les plus aguerris, la solution Wget sous Linux (existant également sous Windows, mais de façon non native) est un programme en ligne de commande qui vous permettra d'obtenir un résultat au format WARC. N'hésitez donc pas à consulter la page du wiki d'Ubuntu, qui offre un premier aperçu de ses fonctionnalités (ainsi que la possibilité d'une interface graphique, pour faciliter son usage) : https://doc.ubuntu-fr.org/wget

Lire les fichiers extraits

Que faire de ces fichiers ainsi extraits ? Comment les lire ? Comment les réutiliser ?

- *Fichier html unique : SingleFile, Save Page We, webrecorder.io*

La lecture d'un tel fichier est simple : les navigateurs web les plus courants n'auront aucun mal à les lire.

- *Fichier html et dossiers en lien : Httrack*

La lecture reste tout aussi simple au travers d'un navigateur web.

L'intérêt de ce format ? Il extrait les différents fichiers d'une page, ce qui peut être intéressant pour sélectionner l'ensemble des images d'un site internet ou d'un de ses sous sites.

Le seul point d'attention à avoir : lors du déplacement des fichiers, il faudra prendre l'ensemble des fichiers sans les séparer.

- *Fichier Warc (Web Archives) : Wget, webrecorder.io*

Pour lire ce type de fichier ou rendre ces fichiers accessibles auprès du public, il vous faudra installer une solution spécifique, soit en ligne soit en local.

Pour une solution en local, il existe le logiciel Webrecorder player.

- *Solution hybride : ArchiveBox*

Globalement, l'interface "lecture" d'ArchiveBox fonctionne très simplement via les navigateurs webs classiques. L'interface donne cependant accès à plusieurs versions, dont deux en html, une en PDF et une prise d'écran ainsi qu'un renvoi vers le site d'origine et vers le Wayback Machine d'Internet Archives. L'intérêt de la démarche est de fournir au futur lecteur plusieurs portes d'entrée : la prise d'écran, notamment, permet de garder l'aspect graphique d'origine de la page, ce que le fichier html ne conserve pas toujours intégralement.

Un des menus permet également le téléchargement d'un fichier au format Warc et Json.