Qu'est-ce qu'une archive du Web ?

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 10 juin 2019

https://www.bortzmeyer.org/archive-du-web.html

Auteur(s) : Francesca Musiani, Camille Paloque-Bergès, Valérie Schafer, Benjamin Thierry
ISBN n°979-10-365-0368-9
Éditeur : OpenEdition Press
Publié en 2019

———————-

Ce très court livre décrit ce qu'est une archive du Web, et les diverses questions que soulève le problème « faut-il conserver tout ce qui a été un jour publié sur le Web et, si oui, comment, notamment compte-tenu de la taille de ces données et de leur rapidité de changement ? ».

Le Web a un peu plus de trente ans et déjà d'innombrables pages Web ont changé voire disparu. Bien des gens seraient intéressés à voir l'état passé du Web : historiens (cf. le précédent livre d'une des auteures, « En construction <https://www.bortzmeyer.org/en-construction.html> »), journalistes (qui voudraient par exemple vérifier le texte qu'un politicien a changé après son élection), simples curieux. . .Mais cela soulève des difficultés techniques et politiques.

Ces difficultés ne sont pas insurmontables : Internet Archive existe et est très utilisé. Ainsi, l'URL vous permettra de voir à quoi ressemblait le site Web de la future AFNIC en juin 1997 (notez comme l'URL est explicite <https://www.bortzmeyer.org/beaux-urls.html>). Et la BNF fait une récolte de tout le Web français <https://www.bnf.fr/fr/archives-de-linternet> (je sais, ce terme n'est pas facile à définir). Ces deux organisations (et plusieurs autres) gèrent un bot qui va ramasser automatiquement les pages, qui seront ensuite stockées. (C'est le même logiciel pour ces deux services, Heritrix <https://github.com/internetarchive/heritrix3/wiki>.) Donc, l'archivage du Web existe mais ce n'est pas facile.

D'abord, voyons les difficultés techniques : le Web est **gros** et grossit en permanence. Il n'existe aucune estimation sérieuse du nombre de pages Web (d'autant plus qu'il n'y a pas de définition claire de ce qu'est une page) mais il ne fait pas de doute que c'est beaucoup. Vouloir stocker tous les états passés de toutes ces pages ne se fait pas avec trois disques durs dans son garage. Mais la principale difficulté technique réside dans la rapidité du changement de ces pages. Certaines pages changent en permanence (la page d'accueil d'un site d'informations, par exemple). Faut-il donc passer toutes les minutes voir cette page ?

Et, ensuite, comment s'assurer que les pages sauvegardées seront encore visibles dans vingt, trente, quarante ans ? Même si on a les données, un site Web en Flash sauvegardé en 2000 sera-t-il encore lisible en 2040 ? Faut-il sauvegarder les données (qu'on ne saura peut-être plus interpréter), ou bien juste une image de la page, rendue par les logiciels existants ?

Un autre problème est celui de la cohérence des pages. Une page Web est constituée de plusieurs élements, par exemple une ressource en HTML, deux en CSS, trois images, et un programme en JavaScript. Toutes ces ressources n'ont pas été récoltées au même moment et peuvent être incohérentes. Les aut·rice·eur·s citent ainsi le cas du site Web du CNRS dont la version « BNF » d'août 2015 montre un bandeau noir lié aux attentats djihadistes de novembre.

Ces difficultés techniques font que l'archivage du Web n'est pas du ressort du bricoleur dans son coin. Il faut de grosses organisations, bien financées, et assurées d'une certaine pérénnité (comme les bibliothèques nationales). Les questions techniques liées à la récolte sont peu mentionnées dans ce livre. Car il y a bien d'autres difficultés, notamment politiques.

D'abord, qui a le droit de récolter ainsi toutes ces pages ? On pourrait se dire qu'elles sont publiques, et qu'il n'y a donc pas de problème. Mais les lois sur la protection des données ne sont pas de cet avis : ce n'est pas parce que quelque chose est public qu'on a le droit de le récolter et de le traiter. Internet Archive considère qu'il est admissible de récolter ces pages publiques, en respectant simplement le `robots.txt`. La BNF s'appuie sur une obligation légale (le dépôt légal est créé par une loi) et ne suit donc pas <https://www.bnf.fr/fr/capture-de-votre-site-web-par-le-robot-de-la-bnf> ce `robots.txt`.

La question peut être sensible dans certains cas. Le livre cite l'exemple des sites Web en `.ao`, récoltés par une organisation portugaise. Bien sûr, ces sites étaient publiquement disponibles et tout le monde pouvait les récolter, mais cela peut être vu ici comme une manifestation de néo-colonialisme tout en sachant que, sans cette récolte de l'ancien colonisateur, rien ne serait récolté.

Ensuite, que peut-on publier de ce qui a été récolté ? Cela soulève des questions liées au droit d'auteur. Pour éviter de froisser les ayant-tous-les-droits, la BNF ne rend pas publique les pages archivées. Internet Archive, par contre, le fait. (Mais l'Internet Archive a déjà retiré des contenus <https://archive.org/post/778/exclusions-from-the-wayback-machine>, par exemple sur ordre de la toute-puissante Scientologie.) Le livre détaille pays par pays les solutions adoptées.

Outre les questions légales liées au droit d'auteur, il peut y avoir des questions éthiques. Par exemple, que penseraient les gens qui avaient contribué à GeoCities si leurs pages de l'époque (publiques, rappelons-le) étaient décortiquées aujourd'hui, alors qu'ils ne s'attendaient pas certainement à ce qu'elles fassent un jour l'objet de tant d'attention.

Et il y a de très nombreuses autres questions à étudier lorsqu'on archive le Web. Bref, un excellent livre, trop court pour tous les sujets à couvrir, mais qui vous fera réfléchir sur une question très riche, ayant plein de conséquences.

Ah, et le livre est disponible gratuitement en EPUB et PDF <https://books.openedition.org/oep/8713>.
