Document generated on 07/14/2025 6:06 p.m.

Documentation et bibliothèques

À la recherche des sites perdus

Eugène Lakinsky

Volume 54, Number 1, January–March 2008

URI: https://id.erudit.org/iderudit/1029251ar
DOI: https://doi.org/10.7202/1029251ar

See table of contents

Publisher(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this document

Lakinsky, E. (2008). À la recherche des sites perdus. *Documentation et bibliothèques*, *54*(1), 31–34. https://doi.org/10.7202/1029251ar

Tous droits réservés © Association pour l'avancement des sciences et des techniques de la documentation (ASTED), 2008

This document is protected by copyright law. Use of the services of Érudit (including reproduction) is subject to its terms and conditions, which can be viewed online.

https://apropos.erudit.org/en/users/policy-on-use/

This article is disseminated and preserved by Érudit.

Érudit is a non-profit inter-university consortium of the Université de Montréal, Université Laval, and the Université du Québec à Montréal. Its mission is to promote and disseminate research.

https://www.erudit.org/en/

Chronique

À la recherche des sites perdus

EUGÈNE LAKINSKY
Service de l'accès à l'information et des ressources documentaires
Ministère de la Santé et des Services sociaux du Québec*
eugenelakinsky@gmail.com

HÉRACLITE D'EPHÈSE a observé, il y a environ quinze siècles, que l'on ne se baigne jamais deux fois dans le même fleuve. Aujourd'hui, la situation est bien pire. Non seulement on ne peut pas se baigner deux fois dans le même fleuve (et dans le cas du Saint-Laurent, même une seule baignade n'est pas conseillée), mais on ne navigue presque jamais deux fois sur le même site Web.

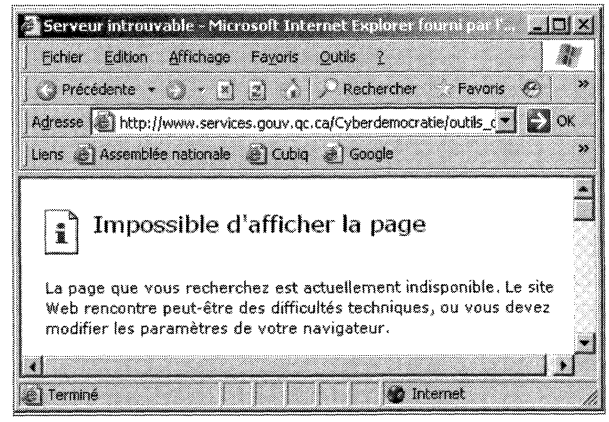
Les sites électroniques changent chaque jour, et parfois même encore plus souvent. Les hyperliens intéressants qui, hier encore, nous menaient vers des pages intéressantes, nous mènent nulle part aujourd'hui. Combien de fois doit-on lire ce triste message : « Serveur introuvable : Impossible d'afficher la page... » Ce qui revient souvent à dire : « *Désolé, la page que vous essayez de consulter n'existe plus* ».

Imaginez que vous avez trouvé un site, disons, parfait : intéressant, facile à parcourir et surtout très pertinent pour votre travail. Vous aimeriez le consulter encore et encore, mais vous avez peur que l'information qui vous intéresse soit archivée ou même détruite, ou que le site déménage ou soit fermé au public.

Il arrive souvent que des organismes publics ou privés effacent des versions électroniques de leurs anciens rapports, communiqués, directives et politiques. Non que les documents soient devenus secrets, mais pour libérer de l'espace disque ou pour alléger la structure du présentoir.

Une page qui disparaît... à jamais ?

Imaginons que vous travaillez dans un centre d'information et qu'il vous faut obtenir des copies de documents de référence portant sur la démocratie électronique. Il y a quelques années, vous aviez préparé une bibliographie sur le sujet et vous y aviez inclus la partie du site Web du ministère des Services gouvernementaux dont l'adresse était :<http://www.services.gouv.qc.ca/Cyberdemocratie/outils_com.html>. On y donnait à ce moment-là accès à plusieurs fichiers PDF de rapports de recherche sur la cyberdémocratie.



Or, après avoir composé la vieille adresse, vous recevez le message suivant : « *Notre site a changé d'adresse. La nouvelle adresse est :* www.msg.gouv.qc.ca. *Veuillez modifier vos signets. Vous serez redirigés vers la nouvelle adresse dans 15 secondes.* »

Après quelques secondes, vous vous retrouvez sur le présentoir www.msg.gouv.qc.ca, plus précisément sur sa page d'accueil. Mais pour retrouver la section qui vous intéresse, il vous faudra recommencer la recherche à zéro.

Ce premier cas est relativement simple : même après le changement de son adresse, le site du ministère des Services gouvernementaux a conservé sa structure initiale. Il suffit donc de modifier légèrement l'adresse <http://www.services.gouv.qc.ca/Cyberdemocratie/outils_com.html> en remplaçant le mot « services » par « msg ». Cela donnera <http://www.msg.gouv.qc.ca/Cyberdemocratie/outils_com.html>, la nouvelle adresse de la section qui vous intéresse.

Mais imaginons qu'il nous faut obtenir le texte du « Compte rendu de conférence : L'objectif à atteindre : la planification de l'analyse d'une consultation en direct » publié par le ministère des Travaux publics du Canada qui, en principe, doit se trouver à l'adresse <http://www.tpsgc.gc.ca/onlineconsultation/text/publications/innovatec_jan_18_2005_report-f.html>. Or, en vous rendant à cette adresse, vous trouvez que « Le site du Centre d'expertise sur la consultation en direct de TPSGC a été archivé ». Cela non plus n'est pas une tragédie. Les pages archivées peuvent être retrouvées dans les Archives du Web du gouvernement du Canada <http://www.collectionscanada.gc.ca/archivesweb/index-f.html>, un nouveau service de Bibliothèque et Archives Canada (BAC). Après quelques secondes de

* L'auteur remercie M. Marc Audet, chargé des projets informatiques à la Bibliothèque de l'Assemblée nationale, qui lui a confié le mandat d'expérimenter le logiciel.

Ce service gratuit met à disposition plus de 85 milliards de pages Web archivées de 1996 à nos jours.

recherche, on retrace le document en question à l'adresse suivante : <http://www.collectionscanada.gc.ca/archivesweb/20060127015514/http://www.tpsgc.gc.ca/onlineconsultation/text/publications/innovatec_jan_18_2005_report-f.html>.

Depuis 2005, la Bibliothèque et Archives Canada (BAC) archive les sites Web du gouvernement du Canada. À l'automne 2007, 100 millions d'objets numériques (plus de 4 téraoctets) de ces données étaient mis à disposition.

Mais que fait-on si une page Web, qui ne faisait pas partie du présentoir fédéral, a été détruite (vraiment détruite) ? Là encore, il n'y a pas lieu de se décourager.

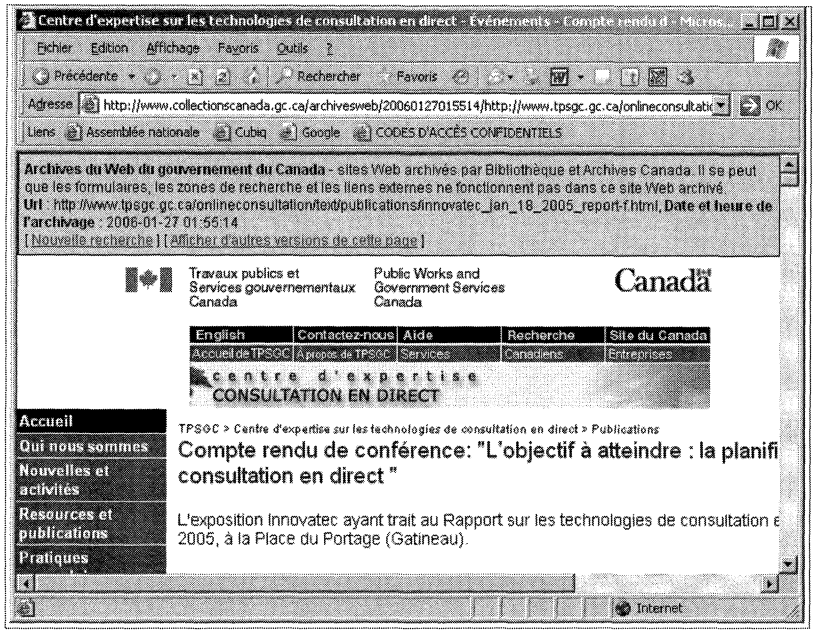
Un bon exemple est une page de Max Nemni, un des co-auteurs du volume « Trudeau, fils du Québec, père du Canada », sur le site Web de l'Université Laval <http://www.pol.ulaval.ca/nemni.html>. Le professeur Max Nemni a pris sa retraite et, en conséquence, sa page personnelle a été détruite. Pourtant, on peut toujours la retrouver avec le site Web de *Internet Archive — Wayback Machine* <http://www.archive.org>. Ce service gratuit met à disposition plus de 85 milliards de pages Web archivées de 1996 à nos jours. On y trouve des sites Web qui n'existent plus ou encore les anciennes versions de nos pages Web préférées. Nous y avons même trouvé l'ancien site personnel d'un ami fait il y a une bonne dizaine d'années et supprimé il y a cinq ou six ans.

Revenons à M. Nemni dont voici la page Web telle quelle se présentait le 6 mai 2003.

On la trouve à l'adresse suivante :

<http://web.archive.org/web/20030506092935/http://www.pol.ulaval.ca/nemni.html>.

De la même façon, on pourra retrouver bien des pages que l'on croyait à jamais disparues.



Que reste-t-il de nos sites Web ?

Mais que faire si un site entier disparaît à jamais ? Un des exemples possibles est celui de la fermeture de Competia <http://www.competia.com>, un site Web sur la veille concurrentielle et stratégique. Competia fut une très bonne source de données en gestion stratégique de l'information. On y publiait d'excellents articles. Il y avait des nouvelles sur les sessions de formation et les séminaires. Un beau jour, la compagnie qui assurait son existence fut achetée par une autre. Le nouveau propriétaire n'était pas intéressé à maintenir le programme. Non seulement a-t-on arrêté les mises à jour, mais même les publications précédentes furent retirées ou gommées.

D'autres exemples ? Au début des années 2000, il existait un excellent site sur le tango argentin, <http://www.todotango.com>. On y trouvait des textes de la plupart des chansons, ainsi que les fichiers musicaux d'enregistrements des années 1920, 1930 et 1940. Tous les classiques du tango s'y trouvaient. Le site était facile à utiliser. Tout était gratuit. On pouvait télécharger les fichiers sur son ordinateur et copier les textes. Hélas, l'âge d'or est terminé. TodoTango ressemble de plus en plus à une boutique virtuelle.

On demande au visiteur d'établir son profil. Les téléchargements ne sont plus permis et il est devenu très difficile de consulter les textes. Maintenant, tout est offert à titre onéreux.

Sur le « sitel » de l'Université de Washington il y a une page du cours « Philosophy 320 : History of Ancient Philosophy » où, dans la section « Texts and Essays », on trouve la collection des articles sur l'histoire de la philosophie <http://faculty.washington.edu/smcohen/320/320Texts.html>. Un vrai trésor pour qui s'intéresse au sujet. Mais qui pourrait garantir que, dans un an ou deux ans, il ne sera pas modifié, déplacé ou effacé ?

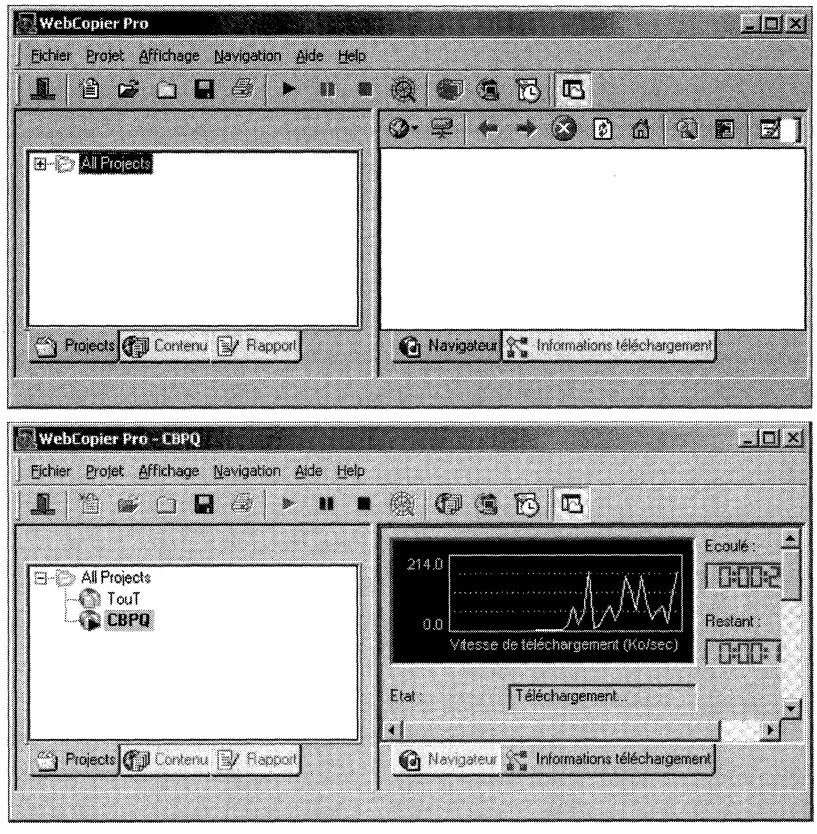
La seule véritable solution de permanence serait de télécharger un site au complet. On pourrait par la suite le garder sur son disque dur, le graver sur un CD ou encore l'enregistrer sur une clé USB.

Pour télécharger un site Web, on peut utiliser un navigateur hors connexion (Offline Browser). Il existe des dizaines de logiciels de ce genre. Sur le fameux site de logiciels à télécharger <http://www.download.com> toute une section leur est consacrée : <http://www.download.com/Offline-Browsers/3150-2377_4-0.html>. Les titres parlent d'eux-mêmes : Website Extractor, SurfOffline, Website Ripper Copier, A1 Website Download, JOC Web Spider, etc. Soit, pas moins de 82 logiciels. Il va sans dire que chacun a ses avantages et ses inconvénients. Une étude comparative prendrait beaucoup de temps et d'énergie.

Or, pour la plupart d'entre nous, il n'y a malheureusement pas d'autre choix que de télécharger un logiciel plus au moins par hasard, en se fiant sur sa description (e.g. : « *Download Web sites and test and convert links into HTTP or local disk* »). Il est à remarquer que beaucoup de ces logiciels offrent une période d'essai gratuit. On peut donc en essayer plusieurs avant de fixer son choix et d'en acheter un. Mais il est clair qu'il n'est pas possible d'expérimenter les 82 titres.

Parmi les navigateurs hors connexion existants, la Bibliothèque de l'Assemblée nationale du Québec a choisi le WebCopier Pro.

De fait, le WebCopier n'est ni pire ni meilleur que ses concurrents. Il fonctionne bien, c'est l'essentiel. Cela dit, s'il faut télécharger un site Web, il le fait rapidement et sans trop d'erreurs. Il ne fait pas des miracles. Parfois, il change les noms des fichiers téléchargés : il n'est pas toujours évident de s'apercevoir que le fichier « Wc2ef12261f959.htm » s'appelait originalement « competition_assessment.asp ». Assez souvent, il faut réajuster le tir, jouer avec les options, corriger des hyperliens.



Cependant, notre expérience d'utilisation de WebCopier Pro permet d'affirmer que, dans la majorité de cas, le téléchargement se fait sans la moindre difficulté. Les problèmes commencent quand certains éléments du site ne se trouvent pas dans le répertoire que l'on télécharge.

Par exemple, admettons que vous avez téléchargé le site du Bulletin *L'intégration* qui se trouve à l'adresse suivante : <http://www.ophq.gouv.qc.ca/lintegration/>. Vous constatez par la suite que les fichiers PDF sont absents. Même le design de la version téléchargée est modifié : il est devenu tout croche et il est difficile d'y naviguer. La raison est simple : les fichiers PDF, les images et les pages de style ne se trouvent pas dans le répertoire « lintegration ». Pour que le site fonctionne, il faut le télécharger au complet : l'URL de la page de démarrage sera donc <http://www.ophq.gouv.qc.ca/>.

Il faut noter également que le volume de certains sites peut être énorme. Aussi, avant de commencer faut-il s'assurer que l'on a assez d'espace libre sur son disque dur.

On peut surveiller le processus de téléchargement grâce à l'interface illustrée ci-dessus.

Le processus de téléchargement peut prendre plusieurs heures, mais s'il dure plus de 24 heures, il faudrait l'arrêter et chercher le problème : jouer avec les options du logiciel, vérifier l'adresse du site ainsi que son « volume ». Peut-être devra-t-on se demander s'il est besoin de télécharger le site au complet ? Un seul répertoire ne suffirait-il pas ?

Puisque la plupart des navigateurs hors connexion fonctionnent d'une façon plus au moins semblable, nous

Notre expérience d'utilisation de WebCopier Pro permet d'affirmer que, dans la majorité de cas, le téléchargement se fait sans la moindre difficulté.

nous permettons de décrire les principales étapes à suivre pour télécharger un site Web à l'aide du WebCopier (voir encadré).

Une fois le site Web[1] téléchargé, on peut y naviguer soit avec le WebCopier lui-même, soit avec son navigateur habituel : Internet Explorer, Firefox ou un autre.

En cliquant sur l'onglet *Rapport*, on peut obtenir la liste des erreurs produites au cours du téléchargement. Mais il ne faut pas paniquer car la plupart d'entre elles ne sont pas graves et n'affectent pas sérieusement la qualité et la fidélité du site téléchargé.

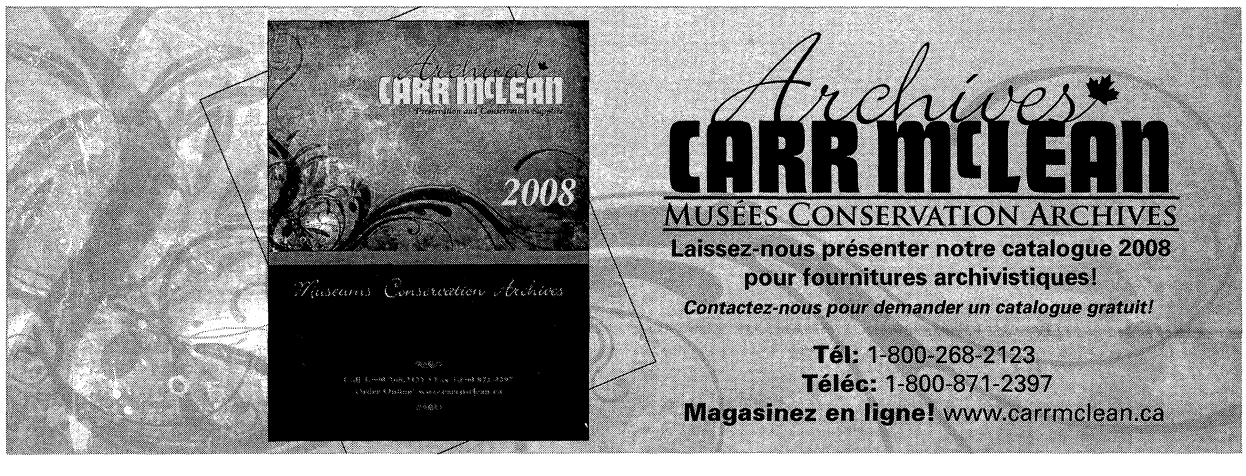
Conclusion

Il n'y a pas de problèmes sans solution. Si une page de site Web qui vous intéresse est détruite, vous pouvez retrouver sa dernière version avec <http://www.archive.org>. S'il s'agit d'un site du gouvernement du Canada, essayez d'abord de le retracer avec <http://www.collectionscanada.gc.ca/archivesweb/index-f.html>.

Mais si un site particulier au complet vous intéresse, que vous le trouvez bien fait et que vous craignez qu'il soit réaménagé, il est probablement prudent de le télécharger et d'en garder une copie. ◉

1. Ici nous prenons comme exemple le site Web de la Corporation des bibliothécaires professionnels du Québec <http://www.cbpq.qc.ca/> que l'on a téléchargé avec le WebCopier.

Petit guide d'utilisation du WebCopier

Étapes à suivre pour télécharger un site Web

1) Cliquer sur l'onglet *Fichier*. Choisir l'option *Créer un Projet*. Cliquer sur le bouton *Suivant*.
2) Entrer le nom d'un site que l'on veut copier. Cliquer sur le bouton *Suivant*.
3) Choisir l'option : *Choisissez l'adresse en entrant son URL*. Cliquer sur le bouton *Suivant*.
4) Entrer l'adresse URL du site à copier. Cliquer sur le bouton *Suivant*.
5) **Ne rien inscrire** dans les champs *Identification* ou *Mot de passe*. Cliquer sur le bouton *Suivant*.
6) Vérifier le cheminement vers un dossier qui abritera la copie du site ainsi que le nom de ce dossier (e.g. : *C :\sites\mon_projet_1*). Cliquer sur le bouton *Suivant*.
7) Choisir l'option : *Spécifier mes propres options*. Cliquer sur le bouton *Suivant*.
8) Choisir l'option : *Dossier de départ & suivant*. Cliquer sur le bouton *Suivant*.
9) Cocher l'option : *Enregistrer le projet et démarrer un téléchargement*. Cliquer sur le bouton *Terminer*.
10) Sauvegarder le fichier (on peut le faire dans le même répertoire qui abritera la copie du site Web, e.g. : *C :\sites\mon_projet_1*).
11) Attendre. Si l'on voit que le téléchargement n'a pas commencé ou que des erreurs graves se produisent, arrêter le processus puis recommencer.
    Note : *Il est normal qu'un téléchargement puisse prendre plusieurs heures.*
12) Une fois le téléchargement terminé, on apercevra le mot *terminer* dans le champ *État*.