5 avaleurs de sites : attention à l'indigestion !

Marchant, Auguste

Publication date:
2000

Département Éducation et Technologie

- Aspirer des sites Web
- Objectifs

5 "avaleurs" de sites

Attention à l'indigestion !

Auguste Marchant

5.71

Mai 2000

Centre pour la Formation à l'Informatique dans le Secondaire

services éducatifs en ligne

Document produit avec le soutien de la DGTRE du Ministère de la Région Wallonne

Projet : Services Éducatifs en Ligne

Document : *Avaleurs.doc*

Point concerné : **SEL 6200 Outils logiciels.**
SEL 3201 Cheminement efficace vers l'information.

Renseignements pratiques :

ECATCH version 3 : http://www.ecatch.com/accueil.htm.

Version 3 en usage libre légèrement bridée. Version 3 complète : 179 FF soit ± 1 500 BEF.

Créé par *Vincent MORELLE* e.mail : bixo@ecatch.com
Téléchargement : http://www.ecatch.com/telechargement.htm

MEMOWEB 98 en version "démo".

Version complète : 299 FF soit ± 1 800 BEF.
prix "éducation" : 300 FF + 150 FF par poste (http://www.memoweb.com/fr/ord4.htm).

Produit de *GOTO Software* (http://www.goto.fr).
Téléchargement : http://www.memoweb.com/fr/trial.htm.

ASPIWEB en version "démo". Version complète : 240 FF soit ± 1 500 BEF.

Produit de *AALWAY Software* (http://www.aalway.net).
Téléchargement : http://www.aalway.net/site/soft/aspiweb/aspi_index_fr.htm .

NAVIGATORCOMPANION en version "démo". Version complète : 1 500 BEF.

Produit de *Eric SARRION* (esarrion@wanadoo.fr).
Téléchargement : http://perso.wanadoo.fr/navcomp/Francais/telech.htm.

WEBCOPIER, en libre accès (1.655 Ko + 13 Ko pour le module français).

Produit de *Maxim KLIMOV*
Téléchargement : http://www.maximumsoft.com/downloads/index.htm.
Caractéristiques : http://www.maximumsoft.com/support/2_0_release_notes.htm
Modules linguistiques : http://www.maximumsoft.com/transl/index.htm.

Avantages et inconvénients

Ces logiciels visent uniquement à "aspirer" les sites Web demandés de manière à pouvoir les consulter "hors ligne". Le dernier permet en outre de sauvegarder les pages consultées, une à une, au moment de leur lecture en ligne.

Les **avantages** de cette technique peuvent au premier abord être manifestes :

- éviter, lors du travail, la lenteur qu'une connexion médiocre peut engendrer dans l'affichage et la lecture en téléchargeant d'avance les pages Web nécessaires ;
- éviter aussi les situations désagréables et ennuyeuses que peuvent provoquer les documents inaccessibles dont l'adresse n'existe plus ou a été modifiée.
- disposer sur sa machine personnelle des pages qu'on est amené à consulter fréquemment.
- limiter, dans le cadre scolaire, le champ d'investigation des élèves au contenu des pages ainsi récupérées. Ce pourra être un gain de temps important tout en continuant à habituer les élèves à la

recherche ; ECATCH, par exemple, dispose d'un module de recherche qui permet de simuler de manière tout à fait réaliste une requête avec un quelconque moteur de recherche.

Ces avantages ne doivent pas cacher certains **inconvénients** ou certaines difficultés.

- Ce pourra être la masse d'informations à stocker sur le disque dur. On pourra y échapper en les gravant sur cédéroms. Si cette option est choisie, il faudra alors graver un nouveau CD chaque fois qu'une mise à jour des pages est faite.
- Ce sera aussi la difficulté de définir précisément le niveau auquel l'aspiration s'arrêtera. Il n'est pas simple de suivre la hiérarchisation des documents lus sur un site. Dès qu'on décide d'en récupérer le contenu , se pose la question : tout le site ? une partie seulement ? que faire des liens vers d'autres sites ? etc.. La réponse est d'importance si l'utilisateur veut éviter l'encombrement provoqué par des documents inutiles. NAVIGATORCOMPANION peut être très utile pour répondre à ces questions puisqu'il est capable d'afficher l'arborescence du site à aspirer.
- La longueur de l'opération n'est pas non plus à négliger. Les plaintes concernant la qualité de la connexion semblent devenir de plus en plus nombreuses. Dans le contexte scolaire, l'utilisateur doit s'attendre à reprendre plusieurs fois le processus suite aux coupures. Et même si les logiciels permettent de reprendre là où l'absorption s'est interrompue, le temps perdu risque d'être assez important. D'autant plus important que, même avec une bonne connexion RNIS par exemple fonctionnant à pleine capacité, les délais restent longs.

Description

L'utilisation de ces cinq logiciels ne pose pas de véritables problèmes. Elle est expliquée en détail dans la suite de ce document.

Deux logiciels, ECATCH et ASPIWEB, sont particulièrement décevants, puisque dans les mêmes conditions de test que les autres, ils n'ont jamais terminé l'aspiration demandée. Ils se bloquent à un certain moment sans raison apparente tandis que la connexion reste ouverte. Par contre, MEMOWEB, NAVIGATORCOMPANION et WEBCOPIER fonctionnent correctement et remplissent complètement la tâche qui leur est assignée.

Tous les logiciels proposent, sous des libellés quelque peu différents, les mêmes paramètres d'aspiration : type de fichiers à ramener, niveaux de profondeurs, suivre les liens externes, etc.. MEMOWEB est le plus complet à cet égard puisqu'il peut récupérer les nouveaux types de fichier, les applets *Java*, les animations VRML et les pages basées sur la technologie *Macromedia Flash*.

Les résultats sont parfois décevants :

- Même si c'est très peu conforme aux normes HTML, beaucoup de concepteurs utilisent dans le code des scripts qui permettent des effets d'images notamment. Or ces scripts font souvent références à des fichiers sans utiliser les balises HTML de référencement. Aucun des aspirateurs testés n'est en mesure d'analyser ces lignes pour en retirer les liens éventuels. En conséquence, il n'est pas rare que des cadres vides se substituent à l'image prévue. À la lecture, est alors assez désagréable.
- Les paramètres établis sont le plus souvent tels qu'un moment donné, les liens ne sont plus suivis. Seul MEMOWEB crée des pages particulières pour indiquer la raison pour laquelle tel ou tel lien n'a pas été suivi ; les autres se contentent de ne pas le suivre et l'utilisateur se retrouve devant les classiques pages d'erreur sans en connaître le motif. Or, pour la qualité de la documentation rapatriée et l'agrément du travail, il importe que l'utilisateur sache la raison du lien "brisé", ne fût-ce que pour pouvoir rectifier le tir en paramétrant mieux son action.
- Plus on évolue vers les technologies "de pointe", moins ces logiciels sont efficaces. MEMOWEB et WEBCOPIER sont capables de rapatrier les applets *Java*, mais seul MEMOWEB est en mesure de récupérer les scènes VRML et les pages construites selon la technologie *Flash*. Il y a donc de plus en plus de risques d'avoir des pages incomplètes ou même pas de page du tout.

Avis

Les "aspirateurs" de site ont certainement leur place dans la panoplie des logiciels à utiliser dans l'enseignement. Bien qu'ils ne soient pas conçus spécifiquement pour cela, ils peuvent y être d'une très grande utilité.

Au delà des dépannages que cette technique peut apporter et des facilités qu'elle offre, elle pose en fait le problème de tout outil informatique : est-ce l'outil adéquat pour obtenir au mieux les résultats souhaités ? Effectivement, les inconvénients liés aux aspirateurs (durée et quantité) sont suffisamment importants pour se poser la question. Inutile d'utiliser un tracteur là où un motoculteur suffit ; l'inverse risque d'être tout aussi pénible.

Reste évidemment la problématique de l'utilisation de ces captures. La qualité de la présentation des résultats laisse à désirer pour tous les logiciels testés sauf MEMOWEB. Là où les autres logiciels "se contentent" de résultats bruts difficilement utilisables tels quels, ce dernier génère des pages explicatives quand un lien n'est pas suivi ainsi qu'une page introductive reprenant les paramètres de la capture. Il crée également un récapitulatif de la capture.

De ce point de vue, pour ceux qui en disposent, ACROBAT 4 semble bien être le plus efficace puisque il permet d'appliquer aux pages capturées toutes les opérations autorisées dans un document PDF. C'est incontestablement le logiciel qui donne le plus d'aisance pour reconstruire un document de bonne facture à partir d'une simple capture. Il est vrai que l'utilisateur dispose, appliqués aux pages Web, de tous les outils utiles à la création de documents PDF.

LES AVALEURS DE SITES

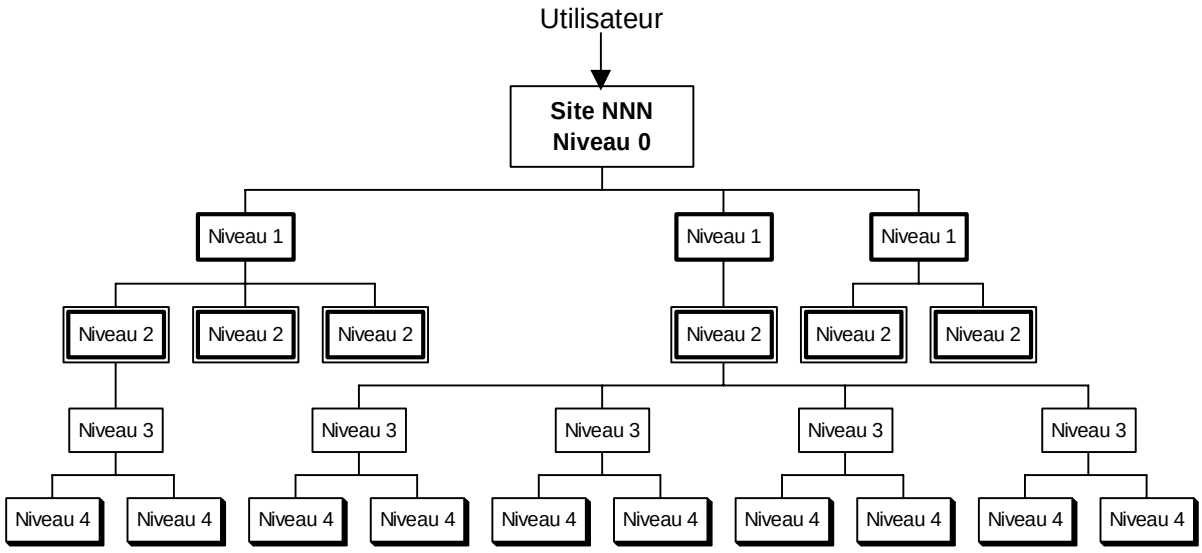
L'argument mis en avant par tous les logiciels de ce genre porte sur la possibilité offerte aux internautes de capturer les pages ou les sites intéressants et donc de les consulter et les utiliser hors ligne. Reste donc à savoir si ces outils remplissent bien leur mission et s'ils le font suffisamment bien et vite pour offrir un avantage consistant par rapport à la consultation directe en ligne. Ce gain doit se situer au niveau de la disponibilité des informations et de la réduction du temps de connexion. Répondraient-ils à ce double critère que la question essentielle resterait pendante : pour quoi faire, dans quel but ? Il existe en effet bien d'autres moyens pour accéder à l'information et la conserver sans être obligé de passer par l'encombrante sauvegarde sur le disque dur de tout ou partie d'un site.

RÉGLAGES PRÉALABLES

Tous les logiciels de ce type répondent aux mêmes normes malgré certaines différences. Tout à fait logiquement, l'utilisateur doit préalablement à toute aspiration, régler certains paramètres. Les premiers sont d'ordre organisationnel :

- bien évidemment l'**URL** du site concerné ; éventuellement, elle sera précisée par le dossier spécifique.
- le **dossier** où les fichiers récupérés seront stockés. Il est loisible à l'utilisateur d'en définir un autre que celui présenté par défaut. Deux cas particuliers :
  - Avec ECATCH, les nouveaux sites doivent obligatoirement être intégrés à des "agents", c'est-à-dire des catégories sous l'étiquette desquelles seront intégrés les sites en question. Cette manière de faire permet ainsi de regrouper sous un même "agent" les pages répondant par exemple à un même sujet. C'est lors de sa création qu'est défini le dossier de stockage. Le logiciel se charge de créer des sous-dossiers pour les différents sites aspirés.
  - NAVIGATORCOMPANION impose ses propres dossiers.
- l'**étiquette** à donner au site qui va être aspiré ; cette dénomination ne correspond pas aux dossiers et sous-dossiers créés par les logiciels sur le disque.

Les suivants sont techniques :



- définition des **niveaux** à explorer. Le niveau "zéro" correspond à la première page du site (ou du dossier). Plus l'exigence sera haute, plus important sera le nombre de fichiers à rapatrier et plus long le temps de connexion. La croissance peut être, selon les sites, quasi géométrique. Ainsi, pour le site du CNDP (http://www.cndp.fr), en travaillant sur le niveau 2, on rapatrie 3,91 Mo, mais 13,1 Mo au niveau 3 et plus de 80 Mo au niveau 5 !
- La **limitation** ou non au seul serveur et/ou seul répertoire indiqué.
  - Ce paramètre ne se trouve ni dans ASPIWEB ni dans NAVIGATORCOMPANION. La limitation porte donc sur le dossier désigné dans l'adresse ou sur tout le site.
  - ECATCH propose simplement les rubriques à cocher "*Sur le même serveur*" et "*Dans le même répertoire*", l'une étant incompatible avec l'autre. Ce choix implique la limitation au seul dossier éventuellement mentionné dans l'adresse ou l'extension à tout le serveur.
  - MEMOWEB et WEBCOPIER se montrent plus précis dans les possibilités offertes :
    - limiter ou non au seul répertoire d'entrée ; ainsi par exemple, pour l'adresse agora-class.fltr.ucl.ac.be/agcl/, faut-il non seulement limiter au seul dossier agcl ?
    - autoriser l'exploration ou non des autres serveurs de même niveau ; par exemple NNN.fltr.ucl.ac.be.
    - suivre ou non les **liens externes** conduisant vers d'autres sites. Si oui, il convient alors de définir sur combien de niveaux.

    En outre, ces deux logiciels disposent de "filtres" qui permettent d'inclure ou d'exclure telle ou telle adresse, tel ou tel dossier. Ces filtres ont priorité sur les réglages précédents. Il est évidemment extrêmement difficile de savoir à l'avance quels sites externes vont être contactés. Mais, lors d'une aspiration, MEMOWEB place dans cette liste, tous les sites externes contactés ; rien n'empêche donc à l'occasion d'une mise à jour, de la vérifier et de l'adapter en fonction de sa plus ou moins grande satisfaction.
- définition des **types** de fichiers à inclure ou exclure ;
  - La proposition habituelle est d'aspirer tous les fichiers sauf les types précisés dans la liste prévue (.ZIP, vidéo, sons, images, programmes, textes).
  - MEMOWEB et WEBCOPIER se montrent plus fins dans la sélection de fichiers à aspirer ou non ; pour chaque type de fichiers (texte, images, sons, …), il est possible de préciser le format (.WAV, .MP3, etc.). L'utilisateur a en outre la faculté de définir de nouveaux types, ce peut être très utile dans un monde où l'évolution et le changement sont permanents.

    N.B. : MEMOWEB se base sur le protocole MIME pour définir le type de fichiers [1]. Deux possibilités sont offertes. Mais, plutôt que d'attendre l'information du serveur avant de télécharger ou non, l'utilisateur peut contraindre le logiciel à prendre l'extension du fichier comme point de repère. La rubrique "*Utiliser l'extension du fichier comme filtre mime-type*" doit alors être activée. Cette seconde solution, en réduisant les échanges avec le serveur, semble plus rapide.

    Pour éviter les limitations que ce système pourrait imposer, l'utilisateur a aussi le loisir de demander la capture de tous les fichiers de type inconnu ou nouveau.

1 MIME : Protocole de communication permettant d'inclure autre chose que du texte dans le courrier électronique, c'est-à-dire des caractères spéciaux, des illustrations, des photos en couleur, des images vidéo ou du son haute-fidélité (http://www.olf.gouv.qc.ca/ressources/internet/index/index.htm).

QUELQUES PARTICULARITÉS

- MEMOWEB et WEBCOPIER autorisent une **limitation** sur le nombre de fichiers à rapatrier, sur leur taille maximale et sur l'espace occupé sur le disque dur. Ces limitations paraissent quelque peu douteuses et difficilement utilisables lorsqu'il s'agit de capturer un site puisqu'on ne connaît pas exactement la quantité de données qu'il représente. Par contre, pour celui dont le disque dur est déjà encombré, … Par défaut, ECATCH se contente de vérifier l'espace libre sur le disque (*Fichier / Préférences générales*). Les deux autres logiciels n'ont rien prévu à cet effet.
- L'**arborescence** du site est respectée par défaut. MEMOWEB permet néanmoins d'accumuler tout en un seul dossier, auquel cas les étiquettes de fichiers sont adaptées.
- Les **liens** sont évidemment adaptés pour une consultation locale. MEMOWEB permet de garder les liens originaux en activant le paramètre "*Rajouter un lien sur l'URL réelle en fin de page*" (fenêtre *Autres options*). L'utilisateur peut ainsi disposer de l'URL réelle de la page et vérifier directement en ligne, par exemple si cette page a été modifiée depuis sa capture.
- La récupération d'un site dépend aussi de la technologie qui y est appliquée. NAVIGATORCOMPANION récupère les fichiers FTP et les places dans un dossier spécial (*C:\ftp*) ; MEMOWEB laisse le choix tant au niveau de la récupération que de l'endroit où les sauvegarder.

  Il est aussi le seul à pouvoir tenir compte des applets Java (.CLASS), des documents 3D VRML (.WRL) et de la technologie *Flash*.
- ASPIWEB propose un assistant pour aider la capture en précisant seulement l'URL ; mais si les réglages personnalisés n'ont pas été définis préalablement, il se basera uniquement sur les paramètres par défaut du logiciel.

  Les réglages, accessibles par le menu *FICHIER / OPTIONS* sont identiques à ceux des autres logiciels. Mais, particularité intéressante, il est possible de filtrer les documents : ne seront capturées que les pages dont le texte contient les critères définis par l'utilisateur.
- Lors de la première utilisation, NAVIGATORCOMPANION analyse le langage utilisé sur la machine et détecte le navigateur installé.
  - Malgré l'analyse, la version mise en place est la version anglaise alors que la version Windows utilisée est la version française. Il faut passer par le bouton *Configuration* (bouton boîte à outils bleue) pour sélectionner le langage adéquat.
  - Le navigateur, *Internet Explorer* en l'occurrence, n'est pas détecté sauf s'il est ouvert à ce moment. Certes, une fenêtre s'ouvre pour signaler la recherche de *Netscape* d'abord, *Internet Explorer* ensuite mais sans le trouver alors que ce dernier a été installé sans modification des paramètres par défaut.

    Il suffit de donner soi-même le chemin vers *Iexplore.exe* pour régler le problème. Ce n'est pas bien gênant … sauf pour des utilisateurs peu au courant de ces notions de chemin et peu soucieux d'avoir en esprit l'organisation des dossiers de Windows.

LES TESTS

Une fois les paramètres définis, on peut lancer le processus d'aspiration proprement dit. Il est à remarquer que ECATCH ne détecte pas une connexion permanente par réseau. Il faut le lui indiquer alors que c'est automatique avec les autres.

C'est à ce moment que l'utilisateur peut s'attendre à certaines surprises et notamment la **quantité de données** récupérées. Les tests ont été faits sur plusieurs sites à des moments différents [(2)]. Pour faciliter les choses, la comparaison s'est faite sur le site du CeFIS. : tout capturer, toute espèce de fichier, mais s'en tenir au seul dossier du CeFIS (http://www.det.fundp.ac.be/cefis/).

- ECATCH : blocage après 2'30" et 8 191 Ko rapatriés. Même résultat au deuxième essai (8025 Ko après 2'25"). L'*Explorateur Windows* renseigne 8,08 Mo sur le disque.

  Les arrêts intempestifs de ce logiciel se sont confirmés dans les autres essais ; bien plus, il a été impossible de terminer correctement et complètement l'aspiration d'un site. Quasiment toujours, le message "**Socket NN fermée**" annonce le blocage. Mais jamais un message clair pour signaler l'interruption. Pire, ces arrêts, ces blocages n'entraînent pas la déconnexion du Web. On imagine aisément les conséquences de pareille situation si, par exemple, une aspiration de site programmée à minuit est ainsi suspendue ; voilà des heures de sommeil qui feront grand plaisir aux opérateurs téléphoniques !

  Le forum (http://www.ecatch.com/forum/) est très significatif de ces difficultés. Les diverses réponses du concepteur mettent en cause les "proxy" par lesquels beaucoup d'abonnés doivent passer et, éventuellement, la présence d'un navigateur déjà ouvert et connecté. Ces deux causes étaient absentes de nos tests. Ce défaut à lui seul est largement suffisant pour le proclamer "éliminé" comme dans certain magazine de consommateurs.

  Il convient ensuite de mettre l'accent sur la lenteur du logiciel : avant de se planter, il avait en 2'30" rapatrié deux fois moins que MEMOWEB en 55" ! Cette lenteur rend encore bien plus vexant tout arrêt de téléchargement.

  > Les déficiences de ECATCH sont telles qu'il est inutile de continuer à en décrire l'utilisation. En son état actuel, la version gratuite téléchargée fin janvier 2000 ne peut engendrer que des ennuis à ceux qui l'utiliseraient. C'est d'autant plus regrettable que la manière de présenter les captures y est assez agréable et offre l'avantage de pouvoir consulter les pages acquises grâce à un navigateur interne et sans être obligé de suivre toute l'arborescence, celle-ci étant schématiquement représentée dans la partie gauche de l'écran.

- Deux difficultés importantes sont apparues à l'utilisation d'ASPIWEB :

  - Que ce soit pour l'aide ou pour la consultation des résultats, un message de l'*Explorateur Windows* signale l'**impossibilité de suivre le chemin indiqué**. Or la notation est parfaitement correcte, les dossiers et les fichiers existent. Impossible donc d'accéder à l'un ou à l'autre sauf en passant par l'*Explorateur* et en ouvrant les fichiers adéquats par ce biais !

    Est-ce à cause de ce problème ? Peut-être, mais la consultation des captures s'est montrée **extrêmement lente**, bien plus lente qu'en se connectant directement au site !

---

2 Agoraclass : http://agoraclass.fltr.ucl.ac.be/
Ciuf : http://www.ciuf.be
Cndp : http://www.cndp.fr.

- Les liens vers les pages qui ne sont pas capturées, quelle qu'en soit la raison, ne sont pas adaptés. C'est ainsi que, régulièrement, là où MEMOWEB crée un page standardisée expliquant pourquoi ce lien n'est pas suivi (limitation volontaire, erreur, …), apparaît la page classique d'erreur : "*Impossible d'ouvrir la page … Impossible de trouver l'erreur DNS ou le serveur dans Internet Explorer*". C'est désagréable avec en plus une impression d'un site mal fait et mal entretenu, alors que, en réalité, c'est le résultat d'une mauvaise capture.

La comparaison des résultats avec ceux de MEMOWEB laisse plus que perplexe. Avec les mêmes réglages (3 et 4 niveaux de capture, tous les fichiers, uniquement le dossier CeFIS du département), nous obtenons :

	Niveaux	***Quantité***	***Délais***	***Liens en erreur***	***Liens ignorés***
MEMOWEB	3	16.2 Mo	1'<	13	75
	4	16.5 Mo	1'	14	64
ASPIWEB	3	6.62 Mo	± 5'	605	
	4	6.62 Mo	± 5'	987	

De manière inexpliquée, ASPIWEB met en erreur quasi toutes les références à des images et ne les capture pas. Résultat : des pages "pleines de vides" désagréables à consulter. Les résultats obtenus sont ainsi faussés et l'utilisateur n'obtient pas ce qu'il souhaite.

> Ce logiciel n'est donc pas à conseiller, même s'il fonctionne légèrement mieux que ECATCH.

- NAVIGATORCOMPANION. La comparaison des résultats avec ceux de MEMOWEB pose problème. Avec les mêmes réglages (3 niveaux de capture, tous les fichiers, uniquement le dossier CeFIS du département), nous obtenons :

	Niveaux	***Quantité***	***Délais***	***Liens en erreur***	***Liens ignorés***
MEMOWEB	3	16.2 Mo	1'<	13	75
NAVIGATOR COMPANION	3	18.4 Mo	± 3'20"	?	

Trois éléments permettent de comprendre la différence de quantité rapatriée :

1. MEMOWEB, dans sa version démo, ne récupère pas les images supérieures à 200 x 200.
2. MEMOWEB signale 10 erreurs FTP et donc ignore les 10 fichiers concernés, soit ± 2.3 Mo alors qu'ils ne posent aucun problème à NAVIGATORCOMPANION. Pourquoi ? La question reste sans réponse actuellement.
3. NAVIGATORCOMPANION respecte moins bien les limites imposées puisqu'il enregistre aussi les premières pages des autres unités qui, avec le CeFIS, composent le département. Il est vrai qu'il n'y a aucun moyen d'imposer cette limitation.

Aucune erreur n'est signalée lors de l'aspiration : les liens brisés sont tout simplement ignorés et l'utilisateur n'en est pas averti. Les 3 pages en erreur (erreur 404) signalées par MEMOWEB ne sont effectivement pas reprises par NAVIGATORCOMPANION exactement comme si les liens vers elles n'existaient pas.

- WEBCOPIER offre les mêmes réglages que les autres logiciels du même type. Il est aussi complet dans son paramétrage que MEMOWEB et, même si la terminologie n'est pas exacte-

ment identique, il n'y a pas de difficultés majeures à procéder à la mise en place d'une aspiration. La comparaison des résultats avec ceux de MEMOWEB est très instructive.

- **Paramétrage** :
  - Cible : http://www.ciuf.be/bibliotheques/.
  - Niveau d'aspiration : 3.
  - Suivre uniquement le dossier de départ et ses sous-dossiers.
  - Exclusion des liens externes.
  - Rapatriement de tous les fichiers.
  - Aucune limitation de taille ni de temps.
- **Résultats** :

	Quantité	*Délais*	*Liens en erreur*	*Fichiers*	*Dossiers*
MEMOWEB	4,89 Mo	1' 45"	35	429	66
WEBCOPIER	4,5 Mo	8'	14	321	186

Les différences de résultats amènent quelques commentaires qui aideront en même temps à mieux comprendre le fonctionnement de l'un et l'autre.

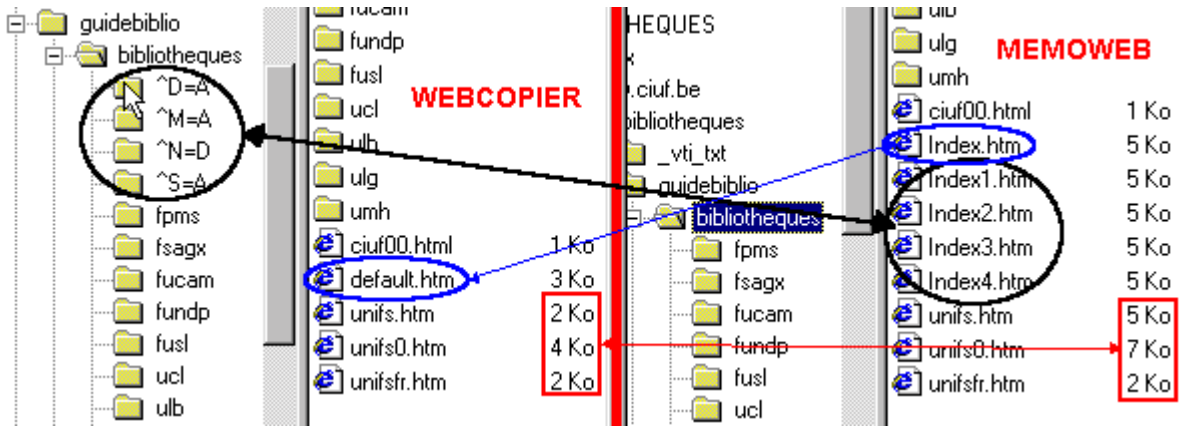
- MEMOWEB annonce en fin de travail une "taille totale utile" 4,89 Mo. En réalité, ce sont quelques 6,76 Mo (taille totale des fichiers et dossiers) qui sont écrits sur le disque.

  En fait, dans son décompte, il ne tient pas compte des pages automatiquement générées pour rappeler les date et paramètres de l'aspiration (*_Accueil.htm* dans la "racine"), expliquer pourquoi un lien n'est pas suivi, lister les erreurs et insérer des pages intercalaires lors de la consultation (version "démo"). Toutes ces pages sont classées dans le dossier *Index* du site aspiré.

  En retirant le contenu de ce dossier, on arrive à 5,44 Mo (429 fichiers et 66 dossiers).

- Même lorsqu'on exige de ne suivre que le seul dossier signalé, MEMOWEB s'écarte de cette exigence lorsqu'il faut récupérer les images accompagnant les pages et classées ailleurs. WEBCOPIER respecte au contraire strictement le critère imposé quitte à n'en rapatrier aucune si ces dernières se trouvent ailleurs ou dans un sous-dossier exclu par le niveau d'aspiration fixé.

  C'est ainsi que, dans l'essai dont il est question ici, aucune image n'est ramenée puisque les concepteurs du site ont placé ces images dans un dossier de même niveau que le dossier *Bibliothèques*.



- WEBCOPIER crée de nouveaux dossiers chaque fois que, suite à une requête, une page dynamique est générée avec la même étiquette. En pareil cas, MEMOWEB regroupe les pages ainsi générées en intégrant une numérotation dans l'étiquette : *Index.htm*, *Index01.htm*, etc..
- En outre les fichiers générés par MEMOWEB sont "plus lourds" ainsi que le montre la capture d'écran ci-dessus.

  L'explication provient de ce que WEBCOPIER se contente de reprendre le code des pages aspirées et de transformer les liens existants en liens relatifs. MEMOWEB par contre, non seulement conserve les anciens liens en leur donnant l'attribut OLDREF (non conforme aux normes HTML), mais en outre introduit du code *JavaScript* lorsque la page incriminée est du dernier niveau d'exploration et donc que les liens éventuels vers d'autres documents qui n'ont pas été aspirés sont coupés. En effet, à ce moment, MEMOWEB génère, grâce à ce code, une page standardisée d'explication. Les conséquences sont triples :

  1. l'utilisateur a toujours l'avantage de comprendre pourquoi tel lien n'a pas été suivi ;
  2. la taille des fichiers augmente ainsi que l'espace disque occupé (± 20% en plus par rapport à WEBCOPIER).
  3. un ancien navigateur pourrait "rechigner" et montrer des "troubles du comportement" face aux attributs non conformes aux règles HTML.

- Quant aux erreurs "HTTP 404" générées par MEMOWEB, elles sont bien plus nombreuses. 15 concernent les images dont il a été question ci-dessus ; 14 sont détectées par les deux logiciels. Certaines sont manifestement des doublons dans lesquels il y a une petite variation "orthographique". Ainsi :

  http://www.ciuf.be/…/repertoire_ressourcesweb/Chimie/titrechimie.html
  http://www.ciuf.be/…/repertoire_ressourcesweb/chimie/titrechimie.html
  ou
  http://www.ciuf.be/bibliotheques/repertoire_ressourcesweb/Droit/titre.html
  http://www.ciuf.be/bibliotheques/repertoire_ressourcesweb/droit/titre.html

  Aucune de ces 4 adresses n'est accessible, mais WEBCOPIER n'en signale qu'une, comme s'il ne tenait pas compte de la distinction majuscules / minuscules. Il ne reste donc que 2 ou 3 cas pour lesquels nous sommes sans réponse ni explication.

CONCLUSIONS DES TESTS

Parmi les 5 logiciels testés, seuls MEMOWEB et NAVIGATORCOMPANION se sont bien tirés d'affaire et aboutissent aux résultats souhaités. Il est par contre bien difficile d'en sélectionner un des trois. Chacun à ses avantages et inconvénients. On retiendra que chacun mène la tâche demandée à bonne fin. Pour le reste, à chaque utilisateur d'apprécier alors en connaissance de cause.

NAVIGATORCOMPANION

- Logiciel réduit à l'écran à une simple barre d'outils flottante (six boutons), sa présence ne gêne absolument pas lorsqu'on est en ligne.
- *Capacité à sauvegarder une à une ou en continu les pages que l'utilisateur examine succesivement.*
- Visualisation préalable de l'arborescence du site à capturer.

MEMOWEB

- *Rapidité d'exécution.*

- *Finesse dans le paramétrage* de ce qu'il faut rapatrier, de ce qu'il faut absolument inclure ou exclure.
- Clarté dans le suivi de l'aspiration.
- Résultats présentés de manière plus "professionnelle" (insertion de pages explicatives pour les liens non suivis). Ce pourrait être du "prêt à l'emploi".
- Capacité d'aspirer des pages produites selon les techniques actuelles (applets Java, VRML, Flash)

WEBCOPIER

- *Mêmes qualités* que le précédent à l'exception de la rapidité et de la mise en forme finale.
- Respect très (trop ?) strict des réglages.
- Un fichier récapitulatif (.log) est généré avec heures de début et de fin de l'aspiration ainsi que les erreurs rencontrées. Ce fichier "texte" est accessible via le logiciel lui-même ou n'importe quel éditeur.

A propos des captures en différé

Tous les logiciels de capture autorisent la capture en différé, c'est-à-dire à un moment où les communications téléphoniques sont moins chères et/ou la connexion meilleure. Il y a évidemment certaines précautions à prendre.

- Veiller à cocher la rubrique *Déconnecter le modem en fin de capture*, afin de couper la ligne lorsque le processus est fini et déconnecter le PC du Web.
- Définir l'heure de début en n'oubliant pas que d'autres travaux sont peut-être, eux aussi programmés (mise à jour à délai fixe par exemple). C'est important pour des logiciels tels que ceux-ci, qui ne supportent pas les multi-sessions.
- Laisser l'ordinateur et le modem actifs.
- Ne pas oublier de donner "login" et mot de passe pour que la connexion au fournisseur d'accès puisse se faire correctement.

Notes sur l'utilisation de MEMOWEB

- Lors de l'aspiration, toutes les données ramenées sur le disque dur défilent à l'écran. Les liens inaccessibles ou les documents écartés en fonction des réglages et des filtres sont signalés. Chaque fois, la raison est succinctement indiquée.

  Lorsque le téléchargement est terminé, le chronomètre (coin inférieur droit de la fenêtre) disparaît et est remplacer par le bouton "*Consulter*". Celui qui dans les "Préférences" aura activé la rubrique "*Appeler le navigateur en fin de capture*", verra alors son navigateur par défaut ouvrir la page d'entrée du site capturé.
- La consultation ne peut se faire que par l'intermédiaire du navigateur installé par défaut sur la machine. Tous les liens étant transformés en liens locaux **relatifs**,. elle se fait exactement comme si on consultait le site en ligne. Seules différences :
  - le logiciel insère une "page de garde" (accueil.htm) reprenant le site et la date de capture ;
  - si, lors de la capture, il a été prévu de garder l'adresse réelle de la page, celle-ci apparaît en bas de page ;

- si un lien mène à une page qui n'a pas été chargée (non trouvée, hors des paramètres définis), le logiciel insère une page explicative. Il suffit alors de revenir en arrière pour reprendre la consultation dans une autre direction ;

- L'exportation des documents vers un autre dossier, un autre ordinateur ou sur cédérom est parfaitement possible puisque toutes les adresses et liens sont relatifs. Il suffit de veiller à copier ou déplacer l'ensemble.

  N.B. : La version téléchargeable de MEMOWEB n'est pas limitée dans le temps. Par contre, elle ne capture pas les images supérieures à 200 sur 200 et insère de la publicité en bas de la fenêtre de capture. En outre, lors de la consultation, une page rappelant qu'il s'agit de la version "démo" est insérée régulièrement toutes les 10 pages.

  La limitation des images peut se révéler assez gênante pour certains sites qui, par leur contenu et/ou technologie mise en place utilisent des images et des animations de plus grande dimension. Néanmoins, ceux qui se contenteraient de cette version peuvent en bonne partie supprimer la publicité en détruisant les fichiers CACHE.DAT, PUB.INI et TEMP.HTM que le logiciel place dans le dossier de la capture.

  Pour supprimer les pages intercalaires, il suffit de supprimer dans le dossier INDEX de la capture tous les fichiers _REG*.HT*.

À propos de NAVIGATORCOMPANION

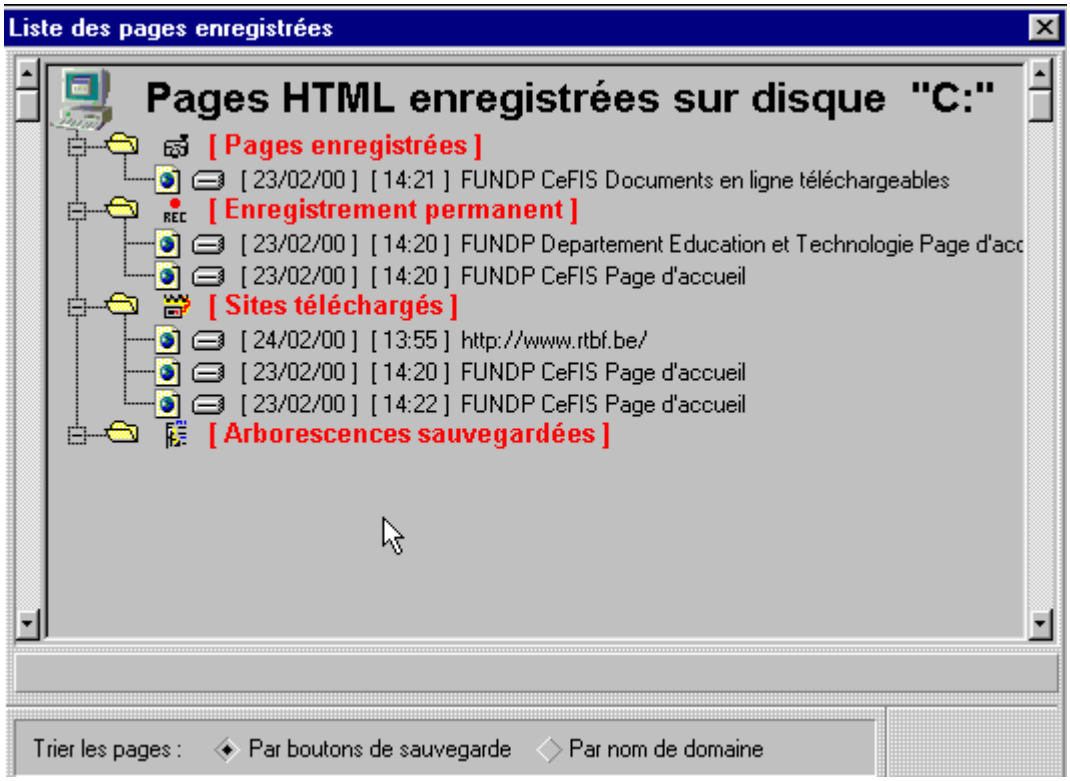
- Seule une barre d'outils flottante signale que NAVIGATORCOMPANION est actif. Elle se compose de 6 boutons :

	Sauvegarder la page chargée (images et texte).
	Enregistrer toutes les pages au fur et à mesure de leur chargement.
	Aspirer un site.
	Visualiser l'arborescence du site, d'accéder à une page précise et, éventuellement, de sauvegarder cette arborescence.
	Boîte à outils : aide, configuration de l'aspiration, langue.
	Afficher l'arborescence des pages aspirées.

- L'ouverture du logiciel entraîne automatiquement celle du navigateur.
- La première sauvegarde ou aspiration entraîne la création d'un dossier *C:\Http* dans lequel est enregistré un fichier *Navcomp.sav* (fichier texte) où sont mémorisées les adresses des pages et sites aspirés. Ces derniers se retrouvent dans des sous dossiers dont l'étiquette est le nom de domaine concerné (*www.det.fundp.ac.be* par exemple).
- La seule aide disponible est celle des quelques pages consacrées au logiciel à l'adresse : http://perso.wanadoo.fr/navcomp/Francais/index.htm.
- La visualisation des résultats commence par la fenêtre *Liste des pages enregistrées* (bouton *ACCÉDER AUX PAGES ENREGISTRÉES*). Les deux modes disponibles, *par session d'enregistrement* et *par nom de domaine*, sont clairs et permettent de sélectionner immédiatement la page ou la séquence de pages à visualiser. Le développement de l'arborescence et les dates d'enregistrement permettent à tout utilisateur de "s'y retrouver" aisément.

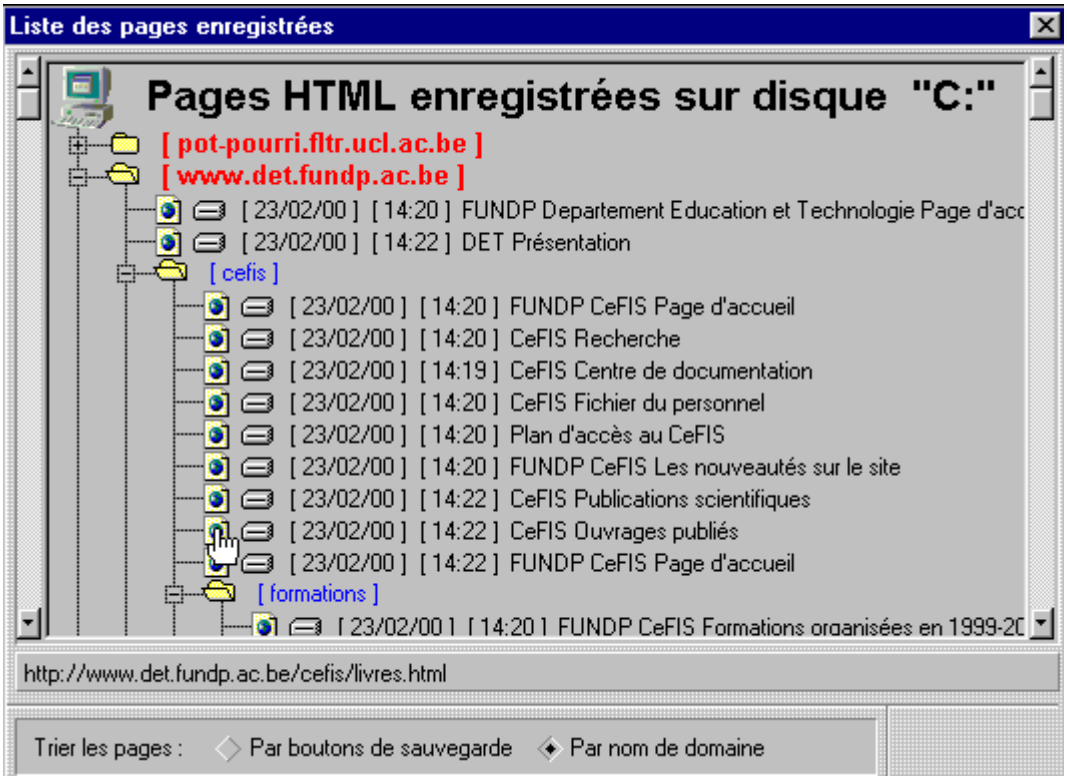
Le second mode paraît le plus intéressant notamment pour des pages acquises en mode *ENREGISTREMENT PERMANENT*. Si l'utilisateur à visité plusieurs pages de plusieurs sites en passant non pas d'un lien à l'autre mais par ses favoris, la visualisation par *session d'enregistrement* ne permettra de voir que la première page de la séquence. Dès lors, il devient impossible à une tierce personne de suivre le parcours du premier utilisateur puisqu'il n'y aura aucun lien entre les pages des différents sites. Il est même probable que l'utilisateur lui-même, avec le temps, en arrive à oublier son parcours.

Par contre, le second mode visualisation permet de développer toute l'arborescence des pages capturées par site (par nom de domaine plus exactement), ce qui est évidemment beaucoup plus clair pour tous. Les captures d'écran ci-dessous valent plus que toute explication.



Ecran 1 : Visualisation par session d'enregistrement

- Lors de la visualisation des résultats, grâce aux menus contextuels, l'utilisateur est en mesure de pratiquer une gestion appropriée des pages. Il peut ainsi les supprimer, les transférer ou les mettre à jour en ouvrant la connexion pour accéder directement à l'adresse répertoriée. Ceci s'applique aux pages et aux domaines.
- Acquisition d'une page par ci par là (bouton *ENREGISTRER LA PAGE AFFICHÉE* ou bouton *ACTIVER L'ENREGISTREMENT PERMANENT*). Une surprise attend ceux qui sont un peu curieux : les pages ne sont pas enregistrées immédiatement et restent dans le "cache" (dossier *Temporary Internet Files*). La vidange (éventuelle) du cache à ce moment entraînera dès lors la perte des informations même si leur référence subsiste dans le fichier de référence *Navcomp.sav*. Elles ne sont sauvegardées qu'à la fermeture du logiciel. Une barre de travail indique alors les fichiers – et leur taille – qui sont transférés dans les dossiers adéquats.
- S'il y a des documents FTP, ils sont placés dans un dossier distinct, *C:\ftp*, chose que MEMOWEB autorise aussi.



Ecran 2 : Visualisation par noms de domaine

À propos de WEBCOPIER

- Par défaut en anglais, mais différents langages, dont le français, sont disponibles. Il suffit de télécharger le module adéquat, de le décompacter et de copier le fichier .WLG adéquat dans le dossier où WEBCOPIER est installé. A l'ouverture du programme, le menu *EDIT/WEBCOPIER OPTIONS* permet de sélectionner le langage voulu.

  N.B. : En fait, contrairement à ce que laissent croire les explications données, le module "français" est intégré au fichier de base téléchargé.

- Toutes les actions de base qu'un tel type de logiciel doit prévoir sont intégrées ici. On notera néanmoins la capacité supplémentaire de filtrer sur base des URL (*OPTIONS DU PROJET / AVANCÉES*). Il est ainsi possible d'exclure ou d'inclure certains dossiers ou documents pour autant que (le début de) leur adresse corresponde au filtre introduit.

QUELQUES CONSTATATIONS

- Le **temps de connexion** varie non seulement en fonction de la quantité de données capturées mais aussi de l'heure à laquelle on se connecte et du serveur auquel on s'adresse. En fait, la vitesse est fonction de divers facteurs parmi lesquels la charge du serveur auquel on se connecte et le "trafic" au moment de la connexion ; Belnet tient constamment à jour des graphiques qui visualisent la densité de ce trafic. Certaines heures sont particulièrement encombrées. L'utilisateur aura donc tout intérêt à bien choisir son moment.

- La récupération des textes, images et autres fichier se base essentiellement sur deux éléments : les liens contenus dans la page et les niveaux d'aspiration tels qu'ils peuvent être définis dans les paramètres. C'est dire que les balises HTML jouent un rôle capital dans le bon rapatriement d'un document. D'où aussi un **défaut** constant quel que soit le logiciel utilisé : certains sigles ou

images sont remplacés par le traditionnel rectangle indiquant que cette image est inaccessible ou inexistante. Ce fait est dû aux lignes de code *Javascript* qu'on trouve très souvent mêlé au HTML et où il peut être fait appel à des images sans code html correspondant.

Ainsi par exemple, dans les pages du CeFIS, le code

```
… <a href="publications/publications.html">
<img src="images/imgPublicationsY.gif"
onmouseOver="this.src='images/imgPublicationsR.gif';"
onmouseOut="this.src='images/imgPublicationsY.gif';" … </a>
```

provoque le téléchargement de l'image "`imgPublicationsY.gif`" mais pas celle "`imgPublicationsR.gif`" qui doit remplacer la précédente au passage du pointeur de la souris.

- Certains serveurs disposent de rubriques qui, sur requête, génèrent des **pages dynamiques**. Cette situation peut amener une quantité de documents bien plus importante que prévu. Le taux de transfert est alors à la baisse, nettement, et les pages chargées pas nécessairement utiles. Est-il par exemple bien utile de rapatrier les archives d'une liste de diffusion bien spécifique, ou sur *Agoraclass* toutes les tables de concordance de vocabulaire latin ? Peut-être, tout dépend du but poursuivi, mais si l'objectif est de récupérer des textes numérisés et libres de droit, …
- L'expérience montre que, en autorisant une capture jusque très profondément dans le site, ou en admettant que le logiciel suive les liens externes, la quantité de données rapatriée grossi très vite, selon une progression quasi **géométrique** (en aspirant le site du *Ciuf* par exemple sur 5 niveaux, ce sont plus de 300 serveurs externes qui sont contactés). Autant le savoir au moment du paramétrage de la capture. À ce moment les performances du logiciel baissent considérablement. L'entrelacement et la multiplication des liens à suivre deviennent tels que le logiciel agit moins vite, peut même "s'y perdre" et donc arrêter la capture.

  Cet effet "boule de neige" pose deux questions dont les réponses induisent un comportement spécifique de l'utilisateur.

  La première est fondamentale : *en quoi ai-je besoin d'aspirer un site ?* Quels motifs m'y poussent ? Quels en sont les objectifs ? Il est fort probable qu'une réflexion sur le sujet amènera souvent à abandonner ce projet au profit de la capture de certaines pages. A ce moment, NAVIGATORCOMPANION se révélera un outil bien précieux en permettant d'enregistrer sans manipulations supplémentaires la ou les pages visitées et uniquement ces dernières.

  Si, par contre, l'utilisateur décide, pour toutes les bonnes raisons qui sont les siennes, de quand même aspirer un site, il lui faudra répondre à la seconde question : *aspirer quoi ?* Le site entier ? Un seul dossier ? Suivre les liens externes ? … Bref, il devra réfléchir à tous les paramètres qu'il imposera. Cette situation devrait, idéalement, impliquer un travail en trois étapes :

  1. Une visite attentive du site afin de se rendre compte aussi exactement que possible de sa structure, de son contenu et de la nécessité ou non de suivre les liens externes qui existent. Cette démarche permettra aussi de préciser le niveau d'aspiration requis.
  2. Une fois l'aspiration exécutée, l'examen des résultats (hors ligne) se révélera intéressant pour déterminer si le niveau d'aspiration est correct, si toutes les pages voulues sont présentes, etc.. Ce sera certainement l'occasion d'affiner le travail et d'exclure ou d'inclure certains dossiers ou documents (in)utiles.
  3. Si les correctifs sont importants, l'utilisateur aura alors tout intérêt à procéder à une mise à jour tenant compte des points précédents.

QUELQUES RÉFLEXIONS

- Tout d'abord, la **qualité** des logiciels. Les essais sont loin d'être exhaustifs et beaucoup diront qu'avec "leur" aspirateur, celui dont ils ont l'habitude, tout va bien. Si c'est vrai, tant mieux. Il n'empêche que les divers tests montrent qu'il faut se méfier des logiciels gratuits ou de certains

shareware. Seul NAVIGATORCOMPANION, parmi les logiciels édités par des "privés", se montre capable de répondre aux besoins de l'utilisateur. Quant aux deux autres, on se demande comment il est possible de mettre "sur le marché" des logiciels qui se bloquent aussi régulièrement.

N.B. : En fait, ce n'est peut-être pas aussi étonnant que cela. Pour rappel, il a fallu une intervention de notre part pour que le concepteur de QUESTY (conception de QCM) corrige une erreur assez monumentale. Son auteur, en toute bonne foi, avait mis son logiciel en téléchargement sur le Web sans se rendre compte que le processus de mélange aléatoire des choix de réponse le rendait totalement inutilisable.

Enfin, pour en terminer avec la qualité, il est bien vrai que NAVIGATORCOMPANION offre des options que même MEMOWEB ne propose pas. Néanmoins ce dernier, par les détails dans la définition des paramètres, les pages additives pour signaler pourquoi un lien n'est pas suivi, la visualisation de l'état d'avancement, montre un niveau nettement plus professionnel. Une combinaison des deux logiciels serait remarquable.

- Tous les essais ont été faits avec la connexion des FUNDP (34 Mbps soit ± 4250Ko). L'aspiration du site du *Ciuf* sur une profondeur de 10 niveaux en ne suivant pas les liens externes mais en capturant tout, a demandé quelques 40 minutes ! Il n'est pas difficile d'imaginer la "monstruosité" du **délai** même avec une ligne RNIS limitée à 64 Ko : plus de 4h30 !

  Il n'est donc pas raisonnable, dans l'enseignement équipé des connexions actuelles, de penser à capturer des quantités importantes de pages et de documents.

- Capturer un site revient à capturer beaucoup de choses : quantité de pages, probablement aussi un certain nombre de documents téléchargeables et certainement une masse assez invraisemblables d'images.

  Sauf motif bien défini, il semble bien inutile de capturer un site complet. Il y a là de quoi encombrer le disque dur d'un tas de choses inutiles. Et si on pense au transfert sur cédérom, ce sera passer beaucoup de temps pour y écrire définitivement des choses dont on ne se servira pas.

  En outre, la matière ainsi récupérée est une matière **brute**. Quelque soit le logiciel utilisé, la seule capture ne suffira pas. Il faudra retravailler les pages récupérées soit pour éliminer les inutiles, soit pour les organiser dans le cadre de l'objectif poursuivi. Ce travail pourrait même être de longue haleine et aller jusqu'à la création de pages de liaison et/ou de synthèse.

Tout ceci amène à se poser la question de la nécessité de capturer un ou des sites. En fait, c'est la question traditionnelle qu'il convient de se poser à propos de n'importe quel outil : pour quoi faire ? Quels sont les objectifs poursuivis ?

Adopter une telle pratique peut se justifier dans deux contextes :

1. La connexion Internet étant **médiocre**, mieux vaut aspirer ce dont on a besoin : la récupération sera lente évidemment, mais elle ne gênera qu'une seule personne, l'enseignant ou l'initiateur de l'activité qui exige les pages et documents ainsi rapatriés.

   Ce faisant, il disposera avec certitude des pages dont il a besoin et n'aura pas à se préoccuper de savoir si la connexion sera bonne au moment voulu. Il ne sera pas soumis aux aléas de l'endroit ou du moment. Le public, quant à lui, disposera des documents requis sans lenteur ni perte de temps.

2. Les pages et documents ainsi récupérés seront utilisés à **plusieurs reprises** dans un travail tel que le même cours à répéter dans plusieurs classes ou une étude de plus longue haleine échelonnée dans le temps.

Outre ces circonstances extérieures, les objectifs poursuivis peuvent être tels que l'utilisation hors ligne de parties de site Web s'impose ; il s'agit alors de ne pas encombrer par une durée trop longue une connexion déjà difficile :

1. Mise à disposition d'une documentation préalablement **triée**. Le professeur, par exemple, ne veut pas que ses élèves se perdent dans le dédale du Web pour rechercher de la documentation. Grâce à l'aspiration de pages, il limite ainsi le champ d'investigation mais aussi la perte de temps.

2. Certaines aspects **techniques** peuvent réclamer une étude relativement longue d'exemples. Qu'il s'agisse d'étudier diverses techniques d'édition (HTML, JAVASCRIPT, etc.) ou certains concepts esthétiques, mieux vaut disposer d'un document "permanent" toujours accessible.

Enfin, il ne faudrait pas oublier qu'il existe d'autres moyens pour garder les informations obtenues. Un agent de recherche tel que COPERNIC permet, lui aussi, d'enregistrer les pages obtenues lors d'une recherche, d'en vérifier le contenu et de mettre à jour les résultats régulièrement, etc.. Certes, ce n'est pas le même outil, mais, une fois de plus, tout est fonction des objectifs poursuivis.

Deux cas particuliers

INTERNET EXPLORER

Les versions récentes d'*Internet Explorer* permettent de sauvegarder certaines pages visitées (menu *FICHIER / ENREGISTRER SOUS*).

Par le menu *FAVORIS / AJOUTER AUX FAVORIS*, les concepteurs permettent de sauvegarder comme favoris l'adresses d'une page et de la rendre accessible hors connexion. En personnalisant cette option, il est même possible de "conserver" les pages vers lesquelles pointent des liens de la première, et ce sur trois niveaux maximum.

Internet Explorer n'en est pas devenu un aspirateur de site pour autant. Il s'agit, une fois de plus, d'un module "croupion" comme il en existe dans beaucoup de logiciels actuellement. Ce n'est pas parce qu'on peut taper une lettre fort présentable dans un tableur que ce tableur est un traitement de texte et *vice versa*. Bref, en parodiant quelque peu : "c'est la couleur d'un avaleur de site, c'est (peut-être) le goût d'un avaleur, mais ce n'est pas un avaleur".

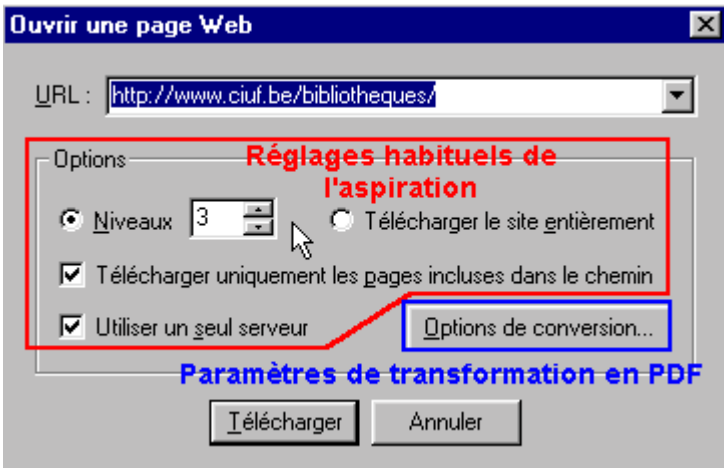
En effet, seules les pages liées au favori créé peuvent être téléchargées. Ces pages forment bloc (dans le dossier *Offline Web Page*) et ne peuvent être examinées qu'à partir de la première qui est répertoriée comme favori. En outre, la suppression du favori entraîne immédiatement la suppression de ces pages qui retournent alors dans le "cache" (dossier *Temporary Internet Files*).

ADOBE ACROBAT 4

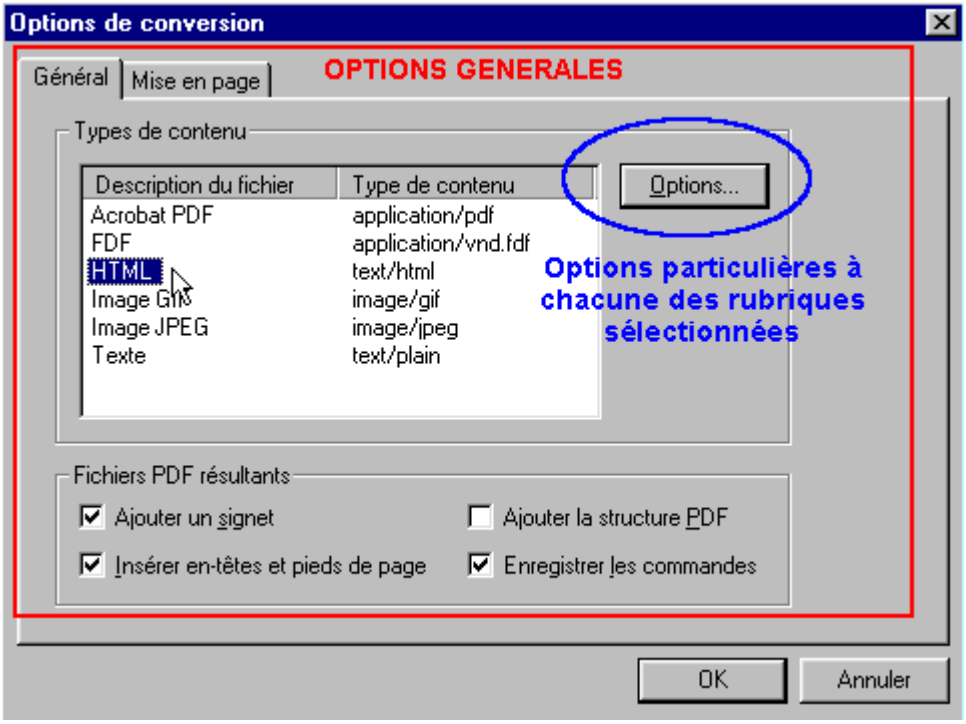
> **Attention !** L'aide (*Téléchargement de pages Web dans Acrobat*) annonce ceci :
>
> Pour pouvoir utiliser Web Capture, vous devez disposer d'***Internet Explorer*** et avoir configuré l'accès au Web dans les propriétés de votre application Internet. Plus précisément, il est essentiel que l'adresse spécifiée pour le proxy dans la zone Serveur du proxy du panneau Connexion soit correcte pour établir des connexions au Web au travers du pare-feu d'une société. Après avoir installé et configuré Internet Explorer, vous pouvez utiliser tout navigateur comme navigateur par défaut. Si votre version d'Internet Explorer ne propose aucune boîte de dialogue des propriétés, vous devez la mettre à jour vers une version récente d'Internet Explorer.

Processus de capture

1. Menu *OUTILS / WEBCAPTURE* ou *FICHIER / OUVRIR UNE PAGE WEB*. Tous les documents et objets sont intégrés dans un fichier PDF.

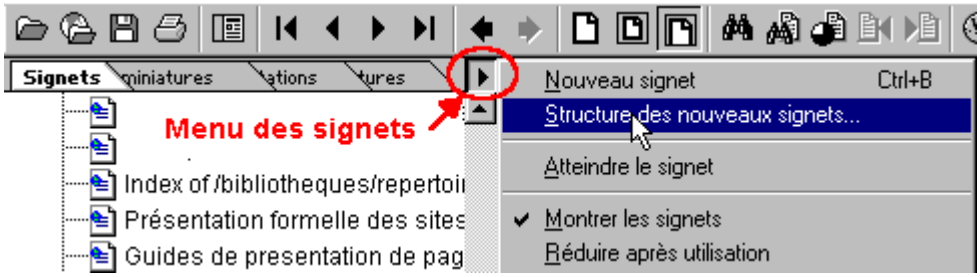


2. Définir les réglages et paramètres de la capture.



- *Général* : pour générer des signets structurés, des en-têtes et pieds de page et une structure PDF pour les pages Web.

  N.B. : A propos de *Ajouter la structure PDF*, l'aide signale que, ce faisant, le logiciel stocke dans le fichier une structure correspondant " à la structure HTML des pages Web originales". Pour y accéder, il faut passer par le menu des signets et sélectionner *Structure des nouveaux signets*.



  Après sélection des balises HTML qu'on désire voir affublées d'un signet, le logiciel les crée mais sans aucune hiérarchie. exactement comme ceux créés au moment de la capture. A première vue donc, cette fonctionnalité paraît peu utile d'autant que la procédure est assez longue en temps et qu'elle fait passer le fichier de 1973 Ko à 3.667 Ko !

  L'option *Enregistrer les commandes* est nécessaire pour "rafraîchir" les pages Web contenues dans le document. ACROBAT constitue alors un nouveau fichier PDF qui répertorie les pages dont les composants ont été modifiés (texte, liens Web, fichiers incorporés et mise en forme). Si le site comporte de nouvelles pages, celles-ci sont téléchargées. Les pages modifiées sont répertoriées dans la palette Signets, sous l'intitulé "Nouvelles pages" et "Pages modifiées".
  (Voir l'aide, *Définition des options de conversion des pages Web*).

- *Mise en page* : pour définir le format des pages, les marges, l'orientation et l'échelle des pages Web insérées dans vos documents PDF.

- *Fichiers HTML* : pour définir les propriétés des polices et autres caractéristiques d'affichage des pages HTML converties au format PDF.

- *Fichiers texte* : pour définir les propriétés des polices et autres caractéristiques d'affichage des pages de texte converties au format PDF.

Résultats : un seul fichier au format PDF de 1973 Ko.

La taille paraît mince par rapport à l'espace disque occupé par la même capture avec MEMOWEB ou WEBCOPIER. Ce n'est pas le site ciblé qui a subi une cure d'amaigrissement, mais, ainsi qu'indiqué ci-dessous, Acrobat ne télécharge aucun document (comment les intégrer dans un format PDF ?). Il ne faut pas non plus négliger la compression que ce format autorise. Elle n'est jamais aussi grande qu'avec des fichiers "texte".

Un rapide examen des pages laisse croire néanmoins que toutes les pages HTML correspondant aux niveaux demandés ont été capturées ainsi que les images associées.

Remarques

- Tout, y compris les images, est intégré immédiatement au fichier PDF.
- Ni les applets *Java* ni les scripts *JavaScript* ni les feuilles de style en cascade ne sont pris en compte. Par contre, les pages construites suivant la technologie *Macromedia Flash* sont capturées. Seuls, les effets visuels que peut produire pareille technique sont "désactivés" même si les liens sous-jacents sont conservés.
- Les documents à télécharger, (protocole FTP ou autre), ne sont pas pris en compte. Bien plus, les liens vers des documents autres que PDF, HTML ou TXT et des images autres que GIF ou JPG sont signalés en erreur.
- L'arborescence du site n'est pas respectée ; toutes les pages sont au même niveau hiérarchique et accessibles soit par les liens traditionnels internes soit par les "signets" qu'ADOBE ACROBAT crée lorsqu'il met au format PDF. Il est néanmoins possible, dans un travail ultérieur, de hiérarchiser soi-même ces signets et d'en créer d'autres.
- Les liens internes sont adaptés ; par contre, les liens externes ou non suivis sont maintenus. Comme rien ne les distinguent des premiers, dès que l'utilisateur clique sur l'un d'eux sans le savoir, le navigateur est activé et propose s'il le faut d'ouvrir la connexion Internet. Ce peut être un inconvénient majeur au moment de l'examen des pages capturées. La commande *OUTILS / LIENS WEB / SUPPRIMER DES LIENS WEB* permet néanmoins de les éliminer.
- Lors de l'intégration de nouvelles pages ou la suppression de pages inutiles, les liens internes sont automatiquement mis à jour (Voir l'aide, *Manipulation de pages Web converties au format PDF*).
- Pour chaque page, l'utilisateur peut en faire afficher tous les liens externes et capturer ainsi les pages qui ne l'auraient pas été vu les réglages imposés (*OUTILS / WEB CAPTURE / AFFICHER LA LISTE DES PAGES LIÉES*). Il ne lui reste plus alors qu'à télécharger les pages qu'il aura sélectionnées.

Avis

- Le module de capture fait partie du progiciel. Le prix d'achat de ce dernier (± 14.000 BEF TVAC, ± 5.300 BEF pour une mise à jour) est évidemment un frein à son utilisation. Encore faudrait-il voir s'il n'existe pas des "prix éducation". Par contre, ce logiciel est devenu quasiment un incontournable de l'édition numérique. Dès lors on peut se demander s'il ne devrait pas faire partie de la panoplie indispensable dans une école.

- La rapidité de capture est quelque peu inférieure à celle de MEMOWEB, mais reste très satisfaisante ; le logiciel doit non seulement télécharger les pages, mais aussi créer les équivalents PDF et tenir compte de la mise en page demandée.
- A première vue, on pourrait reprocher de ne pas garder l'organisation du site capturé. Réflexion faite, ce n'est pas nécessairement négatif. Il est toujours possible de "naviguer" d'une page à l'autre grâce aux liens internes et les signets permettent de visualiser très vite le contenu, un peu comme dans une table des matières.

  Le fait de prendre ce qu'on appelerait si c'était vrai, une photographie de chaque page capturée et d'en faire un seul fichier permet d'éviter tous les difficultés liées à la gestion des dossiers parfois nombreux que peut générer la capture d'un site.
- Par contre, ce qui est particulièrement intéressant, c'est que l'utilisateur dispose de tous les outils d'Acrobat pour retravailler les pages capturées et les adapter strictement à ses objectifs sans avoir à manipuler le code HTML puisque tout est devenu "PDF". Sans vouloir décrire toutes les manipulations autorisées sur un document PDF, on notera :
  - organisation différente des signets (hiérarchisation, suppression, ajout) ;
  - intégration de documents en entier ou en partie issus d'autres sources (Word, Excel, Powerpoint, …) ;
  - insertion de formulaires ;
  - adaptation à un autre format de page que le A4 traditionnel ;
  - modification de l'ordre séquentiel des pages ;
  - suppression des pages inutiles sans avoir à se soucier des liens qui seraient brisés ;
  - ajouts d'effets divers (notes, son, vidéo, …) ;
  - création d'un nouveau document à partir de plusieurs sites capturés ;
  - etc..

  En d'autres termes, d'un point de vue méthodologique, ce module de capture accompagné de toutes les possibilités offertes par ACROBAT, permet à un enseignant de générer un document (qui peut être interactif et multimédia) parfaitement terminé, répondant aux objectifs fixés et entièrement compatible avec les diverses plates-formes qu'on peut trouver.

  Ceux qui disposent de cette version 4 du logiciel, ont manifestement entre les mains, pour les opérations dont il est question ici, un outil de grande classe.
