WEBEARLY 3 et MEMOWEB 3
Des outils pour accélérer la consultation et capturer des sites web

1 - Cédérom WEBEARLY 3 : accélérer la consultation des sites Web

WebEarly 3 est un logiciel qui permet de gagner du temps (et de l'argent) au cours des consultations de documents sur le Web, de retrouver facilement les adresses visitées et toutes les adresses de courrier électronique (e-mail) rencontrées sur les pages Web afin de les exporter ou de les utiliser directement pour un publipostage (mailing) depuis le logiciel, de filtrer sur demande les images externes à caractère publicitaire.

Installation du programme

Elle démarre automatiquement à l'insertion du cédérom dans le lecteur et il suffit de suivre les instructions qui s'affichent à l'écran. Pour que WebEarly fonctionne correctement il faut ensuite inscrire ses coordonnées dans la configuration proxy du navigateur utilisé. Le lancement de WebEarly doit impérativement précéder le lancement du navigateur utilisé, d'ailleurs WebEarly vous propose de se charger lui-même du lancement et de la fermeture du navigateur.

Paramètres et réglages

Le *Temps d'attente* maximum, qui correspond au délai au-delà duquel un essai de connexion n'ayant pas reçu de réponse sera considéré sans succès, doit être défini ainsi que le nombre de tentatives avant échec. Par exemple, si le premier paramètre est fixé à 30 secondes et le deuxième à 3 essais, WebEarly effectuera pour chaque page Web une première tentative de connexion en attendant trente secondes, puis, en cas d'absence de réponse, il fera deux autres essais dans les mêmes conditions. WebEarly fonctionne comme un serveur proxy local par rapport au navigateur utilisé. Pour cette raison, si WebEarly est actif, le navigateur affichera dans sa propre configuration proxy les coordonnées de WebEarly pour le protocole HTTP, à savoir http://localhost. Si

auparavant le navigateur se servait d'un proxy, il est naturel que WebEarly utilise ce même proxy à son tour. Pour ce faire, WebEarly interroge à son premier lancement la configuration proxy du navigateur, avant d'y implanter ses propres coordonnées.

Les limites de l'anticipation

WebEarly met en œuvre une logique d'anticipation tout en laissant la possibilité de la personnaliser et de paramétrer le nombre de pages rapatriées simultanément. La valeur par défaut convient à un accès Internet par modem sur une ligne téléphonique, avec un accès plus rapide on peut augmenter cette valeur. Le *Nombre maximum* de pages à anticiper sert à limiter à court terme le rapatriement anticipé des documents. Au-delà de cette limite, les références les plus anciennes seront oubliées au profit des candidats nouveaux, au fur et à mesure de votre navigation.

La Conservation de la file d'attente

Le pourcentage de *Conservation* de la file d'attente introduit un contrôle supplémentaire de la liste de documents en attente de rapatriement anticipé. Chaque document introduit dans la file d'attente a une priorité qui lui est propre. Si une autre page est demandée avant qu'un document ne soit anticipé, sa priorité sera réduite d'un certain nombre d'unités de valeur, afin que les nouvelles références détectées aient plus d'importance dans le processus d'anticipation. De cette manière, si on navigue d'un site à l'autre sans effectuer de retour en arrière, les premiers documents seront des candidats de plus en plus improbables au rapatriement, car leur priorité n'aura de cesse de baisser.

L'option *Ignorer* les pages non lues au cours des sessions précédentes prend en considération les habitudes de navigation.

La section *Paramètres* pour les raccourcis et les adresses permet de suspendre ou de réactiver la mémorisation de noms raccourcis pour les sites visités, ainsi que la capture d'adresses de courrier électronique rencontrées, avec leur contexte.

Les publicités : WebEarly peut détecter toute image jointe à une page Web, mais externe au site. Un réglage permet de définir à la fois les critères selon lesquels une image externe sera considérée ou non comme étant une publicité, ainsi que la manière de la traiter.

Types de documents anticipés : permet de contrôler quels seront les types de ressources à anticiper ou introduire d'autres types de documents. Une liste déroulante présente systématiquement les derniers types rencontrés pendant la navigation en cours.

La fenêtre *Anticipation* à pour vocation de présenter à la fois la liste des pages rapatriées par WebEarly et le déroulement en temps réel de sa lecture anticipée.

Le menu *Statistiques* de la fenêtre *Anticipation* ouvre l'accès aux informations sur le rendement global de l'anticipation.

Le *Gain de temps* mesure la performance de fourniture de pages anticipées en local par rapport au temps passé pour rapatrier les autres depuis Internet. Par exemple, si le temps d'accès en local depuis le disque dur est 50 fois plus court que le temps moyen de rapatriement, mais que seulement 20% des données ont pu être anticipées, le compteur affichera un gain de 50 x 20% = 1 000%.

La capture d'adresses mail : WebEarly mémorise toutes les adresses de courrier électronique (e-mail) qu'il rencontre dans les pages Web consultées. La fenêtre *Adresses* mail vous présente la liste de toutes les adresses récupérées.

Le menu *Édition* permet d'effectuer une recherche textuelle sur l'ensemble de la liste, en spécifiant une ou plusieurs colonnes concernées (adresses, sites, titres des pages d'origine).

Pour exploiter une sélection d'adresses capturées, le menu *Fichier* offre le choix entre l'exportation (en format texte, texte en colonnes pour tableurs, base de données compatible dBase ou encore page Web) et l'envoi multiple d'un courrier.

Configuration nécessaire

Compatible PC 486 ou supérieur avec Windows 95 ou Windows NT4, 4 Mo de Ram, 3 Mo sur le disque dur Modem, connexion Internet, navigateur Internet Netscape Navigator (version 2.0 ou supérieure) ou Microsoft Internet Explorer (version 3.0 ou supérieure).

2 - CÉDÉROM MEMOWEB 3 : capturer et surfer off-line

MemoWeb est un logiciel, « aspirateur » de sites qui permet de recréer en local sur votre PC tout site Web disponible sur Internet. Il se connecte à votre place, navigue automatiquement dans le Web, cible et stocke sur votre disque tout ce qu'il rencontre : pages d'information, images, sons, vidéos... Il récupère à moindres frais pendant les heures creuses les sites qui vous intéressent d'où un gain de temps et d'argent.

Il est également multitâche : c'est comme si vous ouvriez simultanément plusieurs fenêtres du navigateur sur le même Web, sans jamais afficher la même page dans deux fenêtres !

Comment ça marche ?

Pour créer un web local, vous donnez à MemoWeb des points d'entrée dans ce web, sous la forme de l'adresse Internet d'une page HTML (en général, ce sera la page d'accueil du web à capturer). MemoWeb se connecte au web comme vous le feriez avec votre navigateur, envoie la requête correspondante et rapatrie cette page HTML. Dès la réception complète de celle-ci, il analyse le contenu de la page (source HTML) pour déterminer deux types d'éléments :

- les images ou objets multimédia inclus dans la page, chacun de ces objets est référencé dans la page par son adresse Internet ;
- les liens vers d'autres pages HTML, associés à des zones cliquables à l'écran. Ces liens sont analysés en fonction de critères fournis pour déterminer s'il faut les explorer ou non.

L'ensemble de ces éléments donne lieu alors à de nouvelles requêtes que MemoWeb envoie au serveur Web.

Chaque nouvelle page HTML reçue est traitée selon le même processus. Les autres fichiers (images, sons...) sont simplement stockés sur le disque. La capture du web s'arrête quand il n'y a plus de pages à explorer.

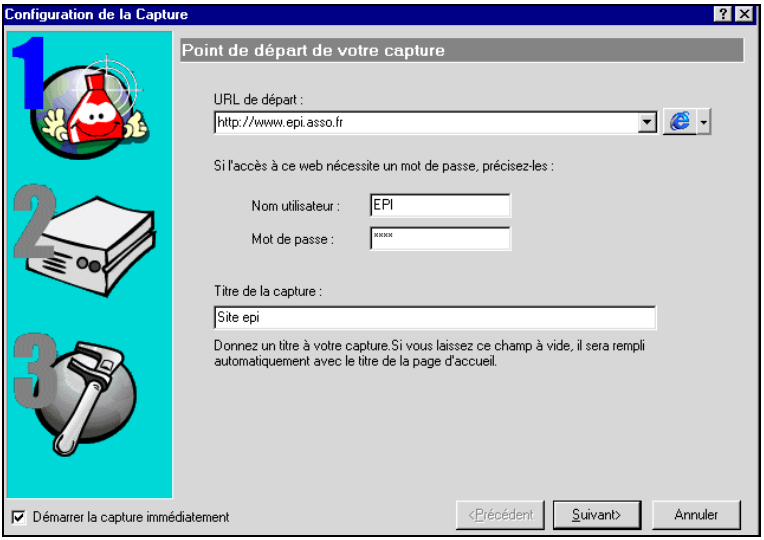
À la fin de la capture, MemoWeb exécute un traitement appelé résolution des liens qui consiste à recharger chacune des pages HTML capturées et à remplacer dans le source les adresses Internet des liens ou des images par les noms des fichiers équivalents capturés.

MemoWeb reconstitue ainsi sur votre disque dur un web complètement autonome dans lequel chaque page HTML pointe vers d'autres pages locales. La dernière phase consiste à créer des pages HTML supplémentaires contenant les index sur les différentes pages et images de la capture. Ces index permettront une navigation plus facile dans le web local.

La capture

Avant tout lancement de capture d'un site Web, vous devez décrire le plus exactement possible les caractéristiques de la tâche que vous confiez à MemoWeb. Cette description repose sur 4 critères principaux :

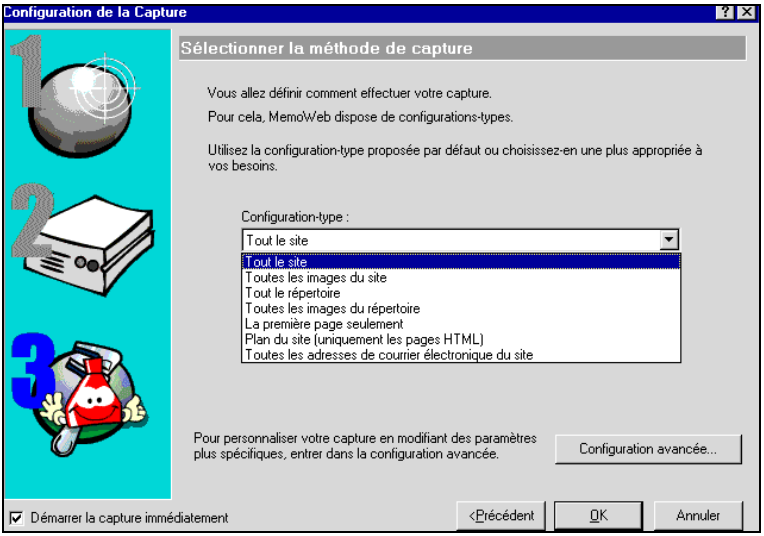
– d'où commencer ? Pour lancer une capture, il faut indiquer à MemoWeb un point d'entrée sur la toile du World Wide Web. C'est l'URL de départ de votre capture. Ce point d'entrée est généralement la page d'accueil d'un web, mais peut être également une page située plus profondément dans l'arborescence d'un serveur. Exemple : http://www.epi.asso.fr/



– où stocker sur votre disque ? MemoWeb enregistre les documents rapatriés dans un répertoire : c'est le web local. C'est ce fichier qu'il faut sélectionner dans la fonction *Ouvrir* pour recharger un web capturé précédemment. Il contient les fichiers capturés, le fichier

structure du web, les pages de compte rendu (répertoire INDEX) et la page d'index général _Start.htm.

– comment et quoi capturer ? Comment correspond aux limites d'exploration que vous allez imposer à MemoWeb et Quoi correspond aux types de documents que vous voulez capturer.



L'outil « recherche »

MemoWeb 3 permet la recherche :

– des pages propriétaires d'un lien (pages contenant ce lien) ;

– des pages HTML selon un texte compris dans le contenu des pages ou dans le titre des pages ;

– des URL selon un filtre au format texte acceptant les caractères spéciaux * et ?. Le caractère * représente tout groupe de caractères jusqu'à celui qui suit le *. Le caractère ? tient lieu de tout caractère individuel. Vous pouvez lancer rapidement une recherche des propriétaires d'un lien à l'aide des menus dans les dossiers *Pages HTML*, *Images*, *Liens ignorés*, etc.

Configuration nécessaire pour fonctionner

Compatible PC avec Windows 95, 98 ou NT4, navigateur et connexion Internet, 32 Mo de RAM, 6 Mo disponibles sur le disque dur (plus l'espace pour les sites capturés).