Dans la série

LES TUTORIELS LIBRES

présentés par le site FRAMASOFT

Capturer des sites avec WinHttrack

Dan

Logiciel : WinHttrack
site : http://www.httrack.com/

Niveau : Débutant

Auteur : Dan site

date de mise en ligne : 11 2002

Framasoft

« Partir de Windows pour découvrir le libre... »

www.framasoft.net

Capturer

des sites avec

WinHttrack

par Dan, novembre 2002

Sommaire

Introduction

Ce tutoriel s'adresse à toute personne désirant capturer un site à l'aide de WinHttrack.

Elle devra avoir quelques connaissances en informatique. Pour capturer un site et pouvoir l'utiliser, il faut maîtriser l'utilisation de l'explorateur Windows.

Tous les sites ne peuvent pas être capturés.

En effet, il faut être en ligne pour de nombreuses bases de données et effectuer des requêtes.

D'autre part les auteurs veulent se protéger de ceux s'approprient des sites.
Pour le faire ils ont à leur disposition de nombreuses méthodes listées, en anglais, dans la documentation de WinHttrack (**abuse FAQ** dont les titres sont ci−dessous).

Même si on peut comprendre les auteurs, la consultation hors ligne marque l'intérêt pour le contenu d'un site ou sa conception. De plus la capture de site peut intéresser des personnes qui ne sont pas mal intentionnées : ***vous par exemple***.

Heureusement, la majorité des concepteurs ne piègent pas leurs sites et se contentent de demander à ceux qui les lisent d'être "corrects".

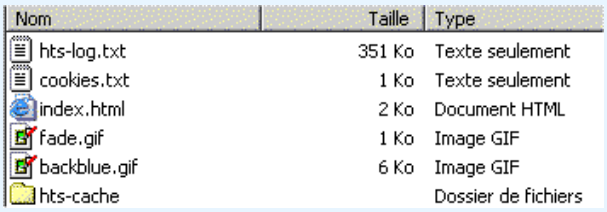
Si vous n'êtes pas découragé, je vais ajouter quelques contraintes puis vous donner quelques clés pour réussir une capture.

Le poste utilisé pour la capture et sa mise au point devra être équipé du même navigateur dans la même version que les postes qui liront cette capture.

Comme la plupart des sites sont prévus pour une lecture avec Internet Explorer (MSIE), c'est celui qu'il vaut mieux utiliser et dans une version récente.

Les versions récentes de Mozilla, Netscape, Phoenix, K−Meleon, Opera et les navigateurs s'appuyant sur MSIE permettent une navigation sur la grande majorité des sites, mais l'affichage d'animations et l'exécution de routines java ou javascript ne sont pas garantis.
Il est aussi souhaitable que les « plugins » Macromedia et que Java soit installés sur tous les postes.
D'autres « plugins », comme IPIX, sont parfois nécessaires.

Quand une capture est incomplète, quelques connaissances supplémentaires en informatique sont indispensables. Parfois il suffit d'ouvrir le fichier hts−log.txt dans le répertoire de la capture,



Nom	Taille	Type
hts-log.txt	351 Ko	Texte seulement
cookies.txt	1 Ko	Texte seulement
index.html	2 Ko	Document HTML
fade.gif	1 Ko	Image GIF
backblue.gif	6 Ko	Image GIF
hts-cache		Dossier de fichiers

parfois il faut aussi maîtriser les possibilités du navigateur, ou encore être capable d'éditer un fichier HTML ou bien avoir une idée de l'organisation d'un site et des fichiers qui le composent.

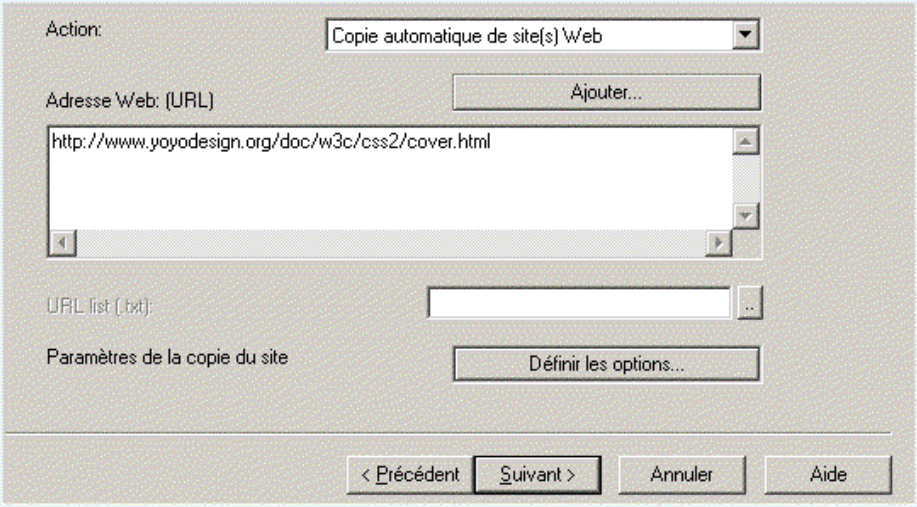
Quand le site est protégé, volontairement ou pas, contre l'aspiration, la maîtrise d'HTML est un minimum.

Une capture facile

Si vous êtes encore là, le tutoriel va traiter des captures aisées dans une première partie puis des captures nécessitant la connaissance d'HTML.

Vous avez installé WinHttrack, sélectionné le français, choisi le répertoire pour les captures (chemin de base) et décidé d'effectuer une capture d'un petit site.

Donnez un nom de projet, par exemple css2 pour la capture des **recommandations CSS2 du W3C en version française** à l'adresse http://www.yoyodesign.org/doc/w3c/css2/cover.html afin de disposer d'une référence pour la feuille de style de votre site.



Mais tout d'abord, quelques remarques à propos des options.

Même si dans ce cas, il suffit de 2 à 3 minutes copier le site, le choix des options dépend beaucoup du type de connexion que vous utilisez.

Si vous utilisez l'ADSL, le câble ou toute autre connexion rapide, inutile dans un premier temps de modifier les options si le site est de petite taille, mais si vous disposez d'un modem 56k, il faut limiter la taille de la capture.

Limiter la taille de la capture

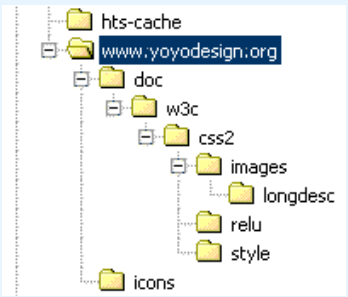
Cliquez sur **Définir les options**.

Dans l'onglet **Règles de filtrage**, j'ajoute **−*.exe −*.zip −*.pdf −*.hqx** et parfois **−*.wav −*.aaif −*.rm** afin d'éviter les gros fichiers, et dans l'onglet **Limites**, **200000** pour **la taille maximale des autres fichiers**.
Je ne coche surtout pas **Noms ISO9660** dans l'onglet **Structure** et je laisse la structure par défaut.
Dans l'onglet **Fouineur**, **Accepter les cookies**, **Analyser les fichiers Java** et **Mise à jour forcée** sont cochés.
Dans **Navigateur Internet**, je vérifie la compatibilité avec le navigateur que j'utilise.

Pour commencer, les autres options par défaut ne doivent pas être modifiées.

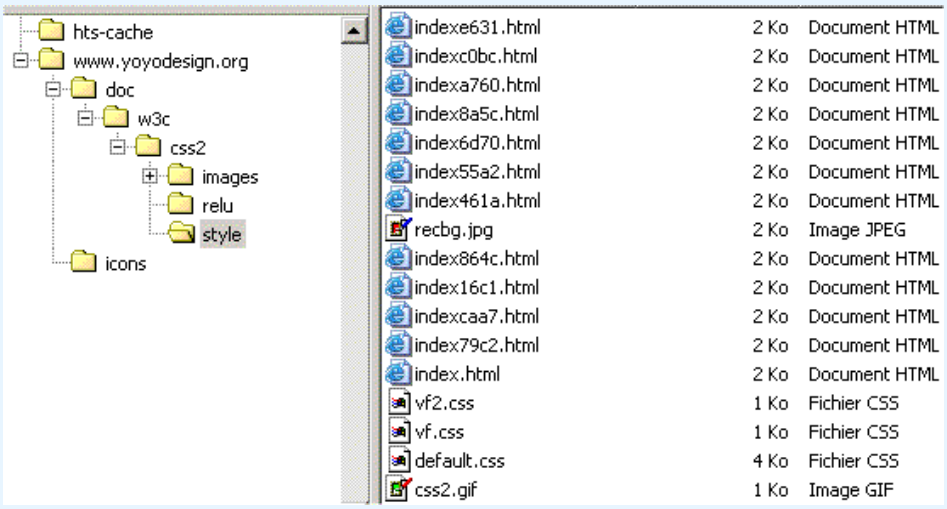
Lancez la capture après avoir précisé votre fournisseur d'accès si nécessaire.

Après quelques minutes, la capture est terminée. Il y a une ou deux erreurs, mais rien de grave.
Vous allez trouver cette arborescence sur le disque dur dans votre chemin de base/css2 :



Le dossier hts−cache contient les fichiers destinés à la mise à jour ou à la poursuite d'une capture, il ne faut pas les modifier.

Les autres dossiers contiennent les fichiers nécessaires au fonctionnement hors−ligne.
Le dossier style contient une partie des fichiers de la capture.



Fichiers indispensables et facultatifs

Vous avez ici quatre types de fichiers indispensables au bon fonctionnement d'une capture : HTML, JPEG, CSS et GIF. Si vous excluez ces types de fichiers dans les **Règles de filtrage**, la capture sera presque toujours incomplète.

Il en existe d'autres : toute la famille des langages qui renvoient des pages HTML (PHP, ASP, CFM...), les fichiers PNG ainsi que les fichiers JS, CLASS, SWF et DIR qui rendent l'exploitation hors−ligne difficile.

Par défaut, ils sont capturés et quand tout va bien, pas besoin de s'en préoccuper.

D'autres fichiers ne sont pas indispensables, mais s'ils vous intéressent, il faut modifier les **Règles de filtrage** et les **Limites** puis relancer la capture à l'aide de **Reprendre une copie interrompue**.

> *D'une manière générale, visitez le site, copiez l'adresse dans le presse−papier pour paramétrer WinHttrack, effectuez une capture avec des limites puis relancez la capture après avoir fait le tour de ces fichiers qui vous intéressent.*
> *Recommencez avec les liens que vous voulez ajouter.*
> *WinHttrack créera dans le répertoire hts−cache des fichiers old.lst, old.ndx, old.dat et old.txt qui peuvent permettre un retour en arrière si le nouveau paramétrage ne donne pas le résultat escompté (supprimez les fichiers new.lst, new.ndx, new.dat et new.txt et remplacez old par new dans les fichiers restants).*

Pour trouver les noms et extensions des fichiers qui vous intéressent, deux solutions :

1. La plus simple comme pour ce site, c'est de passer le curseur sur le lien (ou clic droit sur le lien et **copier le raccourci)** et noter le nom du fichier qu'on souhaite capturer et qui apparaît dans la barre de statut.
   Dans le fichier refs.html de la capture, vous souhaitez le fichier PDF. Ici, l'adresse vous est donnée, sinon vous la trouvez à côte du .

   Ici, le fichier est extérieur au site. Dans ce cas, il suffit d'ajouter la ligne **+ftp://sgigate.sgi.com/pub/icc/CC32.pdf** (ou + et clic droit **coller**) dans les **Règles de filtrage**, ôter la limite de **200k** dans **Limites** puisque le site n'est pas gros et relancer la capture.
   Quand le fichier est dans le site, on peut enlever l'option **−*.pdf**, mais tous ces fichiers seront alors téléchargés.
2. Quand le nom n'apparaît pas dans la barre de statut, ou si on cherche la difficulté, on ouvre le fichier (ici refs.html) avec son éditeur favori, on recherche les fichiers (ici CC32.pdf) dans la page ou le texte qui est affiché à proximité du fichier intéressant. On peut alors noter les noms des fichiers à télécharger et modifier les **Règles de filtrage** ou bien télécharger le(s) fichier(s) avec un utilitaire, le(s) copier dans la capture et modifier les liens dans le fichier HTML.

Ajouter des liens

Quand le nom de fichier qui s'affiche dans la barre de statut est un fichier HTML ou un nom de dossier, on ajoute un lien vers une autre partie de site ou vers un autre site.

Si vous souhaitez ajouter http://www.yoyodesign.org/doc/w3c/w3c.html qui se trouve plus haut dans l'arborescence du site et qui n'a donc pas été capturé, ajoutez **+http://www.yoyodesign.org/doc/w3c/w3c.html** dans les **Règles de filtrage,** mais vous risquez de désigner tout ce qui se trouve sous http://www.yoyodesign.org/doc/w3c/ et cela peut représenter des heures de capture à moins de compliquer les règles.

Si vous souhaitez ajouter http://www.w3.org/Consortium/Translation/French/, faites de même, mais prévoyez plusieurs mégaoctets de capture.

Si vous oubliez la barre en fin du nom de dossier, c'est un répertoire de plus et tous ses sous répertoires qui seront aspirés !

Ça se complique

Vous avez vu dans la capture précédente comment paramétrer une capture, limiter le nombre de fichiers téléchargés, ajouter les fichiers intéressants ou ajouter un site ou une partie de site.
Vous savez que les fichiers HTML, ASP, PHP et CFM que WinHttrack sauve sur le disque dur avec l'extension html compliquent le paramétrage car ils ajoutent à la capture tous les fichiers inclus dans les **Règles de capture** qui composent la page.
Vous avez noté qu'on doit visiter quelques pages du site et les pages qu'on veut ajouter pour avoir une idée de ce qu'on va capturer.

Néanmoins, comme la majorité des captures elle ne pose pas de problème.

Vous pouvez graver le répertoire même si vous n'avez pas coché l'option **Noms ISO9660** après avoir supprimé dans le sous répertoire hts−cache les fichiers old.* qui ne servent plus à rien.

Quelques captures sont plus difficiles à réaliser. En voici un exemple.

Nous allons effectuer la copie d'un site − La ferme aux crocodiles − qui pose plusieurs problèmes : les applets java et les fichiers Flash.
Visitez la page suivante : La ferme aux crocodiles.
Installez le "plugin" Flash si votre navigateur n'affiche pas le crocodile en milieu de page.

Lancez WinHttrack.
Entrez **http://www.lafermeauxcrocodiles.com/** dans l'**Adresse Web**.

Dans **définir les options**, assurez vous que **Noms ISO9660** n'est pas coché. En effet, les applets java font appel entre autres à des fichiers dont l'extension est CLASS, soit une lettre de trop. Ils seront sauvés avec une extension CLA et ne pourront donc pas être interprétés par le "plugin" Java. Cette remarque est valable avec toutes les extensions de plus de trois lettres sauf HTML (et encore, pas toujours).

Lancez la capture. Elle dure environ 25 minutes avec un modem 56k. Il n'y a pas d'erreur dans le compte rendu de capture.

Explorez la copie du site.

L'intro, écrite en Flash, fonctionne.
La page d'accueil qui apparaît ensuite est incomplète :

1. Il manque plusieurs images, sauf si vous avez visité le site et chargé la totalité de la page.
2. Un cadre gris apparaît à droite de l'écran.

Les images manquantes en haut de l'écran font appel à un script VB pour afficher de la publicité.
A mon avis il n'est pas utile de les télécharger.
Si vous voulez vraiment le faire, modifiez les **Règles de capture** ou utilisez le cache d'Internet Explorer (nous verrons plus loin comment l'utiliser).

Le cadre gris est typique d'un applet java. Quand vous passez la souris, "Applet démarré" ou un message d'erreur s'affiche dans la barre de statut.

Pour régler ce genre de problème, il faut afficher la source.

Comme ce site utilise des cadres (frames), le plus simple c'est de télécharger et installer les accessoires pour IE5 en français. Ils sont disponibles à ce jour à cette adresse : http://loranger.free.fr/webtools/files/ie5wafr.exe.
Une fois installés, ils permettent d'ouvrir le cadre où se trouve le curseur de la souris dans une nouvelle fenêtre à l'aide d'un clic droit.

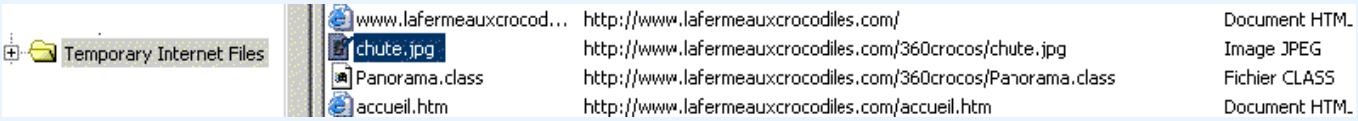
Sur la page d'accueil en français, clic droit sous le cadre gris, et la page pageacueill.htm s'ouvre.
Affichez la source (**Affichage** puis **source** dans MSIE) ou ouvrez la page avec votre éditeur HTML.
Vous allez trouver le code suivant.

```
<td><applet codebase="360crocos/" code="Panorama.class"
                        width="340" height="117" align="middle">
<param name="picture" value="chute.jpg">
<param name="delay" value="50">
<param name="y_add" value="1">
</applet></td>
```

Dans le répertoire www.lafermeauxcrocodiles.com/360crocos on trouve bien Panorama.class mais pas chute.jpg.
Il est donc normal que l'applet n'affiche rien.

Pour régler le problème, il faut visiter avec MSIE la page qui pose problème et laisser le téléchargement de tous ses éléments se terminer.
Il faut ensuite trouver le cache de MSIE. Il s'appelle Temporary Internet Files et son emplacement dépend de la version de Windows et de l'utilisateur :
par exemple C:/Documents and Settings/vous/Temporary Internet Files ou encore C:/Windows/Temporary Internet Files.
Cherchez le cache et classez les fichiers par adresse Internet.

Dans la liste cherchez www.lafermeauxcrocodiles.com. Vous allez trouver :

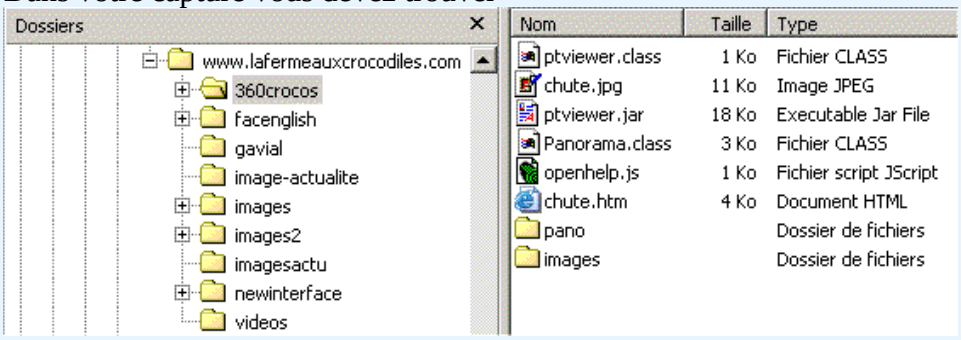


Vous pouvez constater ici que Panorama.class n'appelle aucun fichier CLASS ou JAR −qu'il aurait fallu copier dans les répertoires de la capture− et que le fichier chute.jpg utilisé est celui du répertoire 360crocos (et pas de 360crocos/pano).

Copiez donc ce fichier dans la capture (dans le répertoire www.lafermeauxcrocodiles.com/360crocos). Windows copie les fichiers du cache en ajoutant [un chiffre], ici chute[1].jpg si vous n'avez pas cliqué sur la visite qui utilise 360crocos/pano/chute.jpg −vous auriez alors chute[1].jpg, chute[2].jpg en fonction du répertoire d'origine.

Renommez chute[1].jpg en chute.jpg.

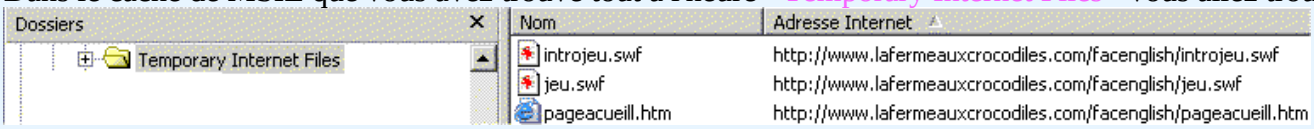
Dans votre capture vous devez trouver



et l'affichage de l'animation sur la page d'accueil doit se faire.
Par chance, La page d'accueil en anglais fonctionne elle aussi maintenant.

Dans la rubrique Jeux (ou Games), seule l'introduction a été téléchargée.
Comme précédemment, il faut visiter le site avec MSIE, lancer le jeu, cliquer sur GO à la fin de l'introduction puis attendre que le jeu soit chargé.

Dans le cache de MSIE que vous avez trouvé tout à l'heure −Temporary Internet Files− vous allez trouver

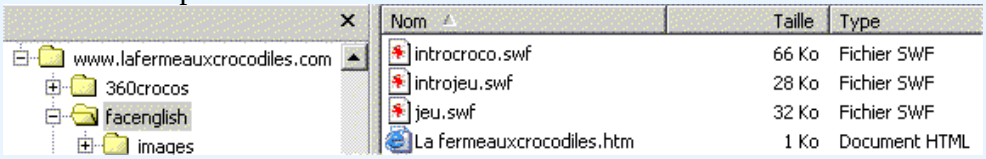


copiez le fichier jeu.swf dans le répertoire www.lafermeauxcrocodiles.com/facenglish/ de votre capture.

Comme à chaque fois, Windows copie les fichiers du cache en ajoutant [un chiffre], ici jeu[1].swf.

Renommez jeu[1].swf en jeu.swf.

Dans votre capture vous devez trouver



et le jeu fonctionne maintenant.

Revenons aux applets java

De retour sur la page d'accueil, pourquoi ne pas visiter la ferme ?

Clic sur le lien, la cascade apparaît, et on nous propose de choisir d'autres photographies.
Clic sur une photo, par exemple la plage. Pas de chance : page vierge.

Pour régler ce genre de problème, il faut afficher la source.
L'applet java dans la page www.lafermeauxcrocodiles.com/360crocos/chute.htm

```
<applet code="ptviewer.class" name="ptviewer" archive="ptviewer.jar"
                width="500" height="270" mayscript="true">
<param name="file" value="pano/chute.jpg">
<param name="wait" value="images/frame/frame_wait.gif">
[...] </applet>
```

charge plusieurs images et gère un menu. Il appelle les pages HTML qui affichent les autres photos.

Dans ce cas aussi, il faut visiter avec MSIE la page qui pose problème et laisser le téléchargement de tous ses éléments se terminer.
Demandez à voir toutes les photographies.
Vous pouvez ensuite tenter deux méthodes :

1. Notez dans la barre d'adresse les noms des fichiers qui sont appelés :
   - www.lafermeauxcrocodiles.com/360crocos/lac.htm
   - www.lafermeauxcrocodiles.com/360crocos/labo.htm
   - www.lafermeauxcrocodiles.com/360crocos/fougere.htm
   - www.lafermeauxcrocodiles.com/360crocos/plage.htm
   - www.lafermeauxcrocodiles.com/360crocos/pedago.htm
   - www.lafermeauxcrocodiles.com/360crocos/grotte.htm.

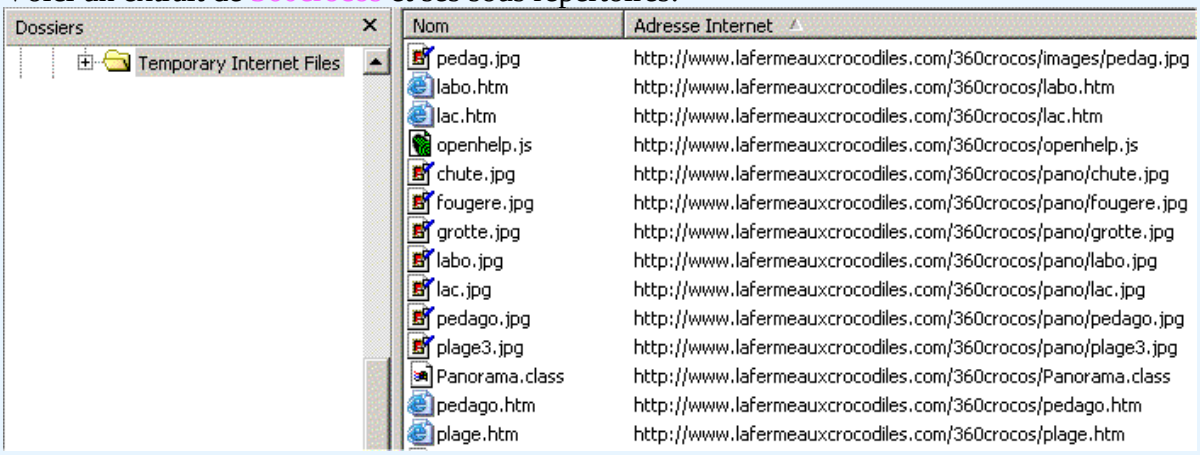
   Ajoutez les dans les **adresses WEB** à la suite de http://www.lafermeauxcrocodiles.com/accueil.htm et sélectionnez **Reprendre une copie interrompue** (si on les ajoute dans les **Règles de capture**, elles ne sont pas capturées).

   Pour cette capture, le résultat est concluant. Toutes les photos s'affichent.

Si cela ne marche pas, utilisez la méthode suivante qui demande beaucoup plus de temps mais qui permet de cerner tous les problèmes.

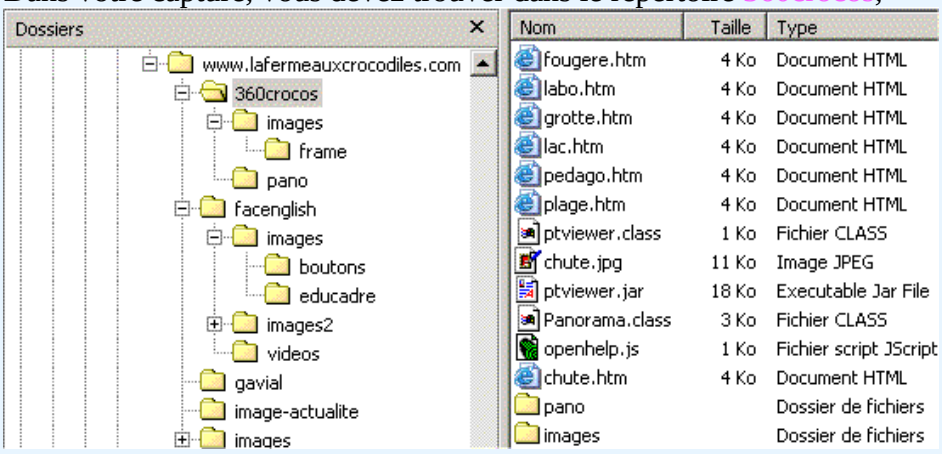
2. Explorez le cache de MSIE −Temporary Internet Files− que vous avez utilisé précédemment.

Classez les fichiers par adresse Internet.
Voici un extrait de 360crocos et ses sous répertoires:

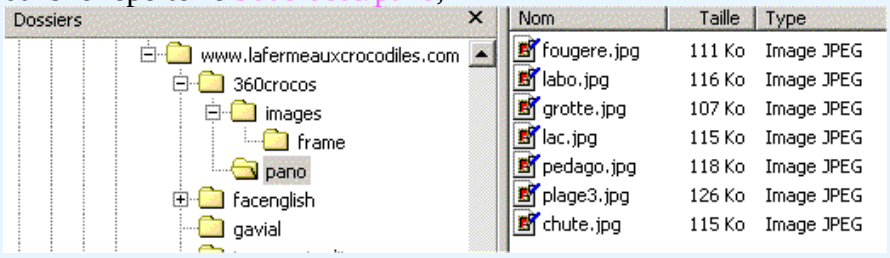


Parcourez tout ce qui concerne le site et copiez répertoire par répertoire tous les fichiers qui manquent dans la capture.
Comme à chaque fois, Windows copie les fichiers du cache en ajoutant [un chiffre].
Enlevez le chiffre entre crochets.

Dans votre capture, vous devez trouver dans le répertoire 360crocos,



dans le répertoire 360crocos/pano,



et ainsi de suite.

Maintenant le menu fonctionne et la visite peut s'effectuer.

Conclusion

Si vous avez tenté les deux captures présentées, vous pouvez aspirer une grande majorité des sites présentant un intérêt car vous avez vu comment paramétrer une capture, limiter le nombre de fichiers téléchargés, ajouter les fichiers intéressants ou ajouter un site ou une partie de site, utiliser la source d'un fichier HTML, utiliser le cache d'Internet Explorer.
Vous pouvez employer les mêmes méthodes pour des problèmes plus complexes : ajout de lien dans le cas d'une redirection, applet appelant d'autres applets, modification des paramètres d'applets et menus écrits en Flash.
Si vous visitez un site avant sa capture et utilisez l'URL trouvé dans la barre d'adresse du navigateur, vous aurez peu de problèmes et éviterez les redirections des sites ayant changé de prestataire.

Il vous restera ensuite à apprendre à modifier une page HTML pour corriger les erreurs dans les liens et dans les noms de fichiers en particulier pour les sites en ASP, PHP et CFM.
Il sera également utile d'interpréter les informations du fichier hts−log.txt.
Il faudra aussi se pencher sur la recherche des fichiers DIR et les liens vers les fichiers son ou vidéo.

Et il restera javascript !

C'est ce langage qui bloque le plus souvent les captures, généralement sans que les auteurs ne le souhaitent.

Pour ce qui est des sites piégés, ce sont les moins intéressants, ils nécessitent toutes les compétences ci−dessus.

Bon courage.

Vous pourrez trouver quelques exemples de problèmes plus difficiles à régler sur le site danzcontrib.free.fr et des réponses sur le forum d'HTTrack.

Vous pouvez signaler toute erreur, faute ou imprécision ici.