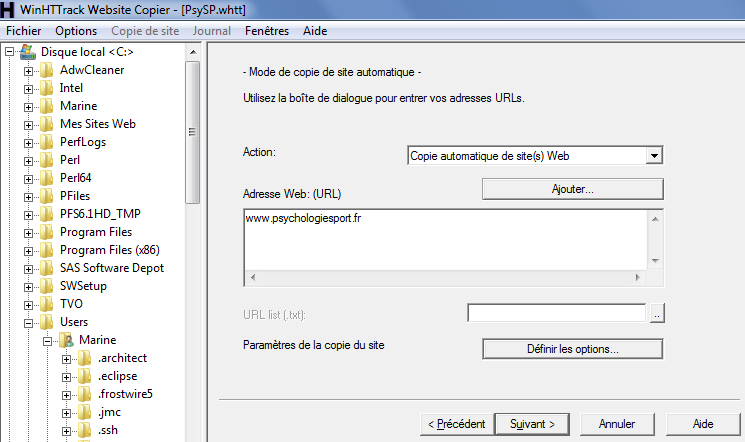
Au cours de nos recherches, on a trouvé un site internet (www.psychologiesport.fr) intéressant pour notre domaine. Ce site est composé d’une centaine d’articles concernant la psychologie dans le sport. On a remarqué en naviguant sur le site que certains articles étaient des liens vers d’autres sites, par exemple vers www.20minutes.fr ou www.lequipe.fr. On souhaitait ainsi récupérer les articles intéressant sur le site psychologiesport.fr ou d’un site indiqué dans un lien.



Figure . Capture écran d'un article de www.psychologiesport.fr dont on veut récupérer le lien vers un autre site

Pour résoudre ce problème, on a téléchargé toutes les pages de www.psychologiesport.fr à l’aide du logiciel WinHTTRACK WEBSITE COPIER.



Nous avons choisi de ne récupérer que les pages du site en indiquant que la limite de profondeur externe est de 0. Ainsi HTTRACK n’est pas allé sur les sites internes autres que www.[psychologiesport.fr](http://www.psychologiesport.fr). HTTRACK organise les pages en utilisant la même structure que le site internet. Donc nos articles étaient répartis dans des dossiers différents. Nous avons codé un programme Perl (lebon.pl) qui déplacer toutes les pages et les renommait car elles portaient toutes le même nom (index.html).

On a ainsi obtenu les pages de ce site mais pas les liens vers les autres sites. Grâce à un autre programme Perl (url.pl), nous avons récupérer toutes les URL qui se trouvaient dans le code HTML des pages en enlevant les liens vers certaines pages qui ne nous intéressaient pas et qui se trouvaient sur toutes les pages de psychologiesport.fr.  
 nous avons fait tourner une nouvelle fois HTTRACK avec cette liste d’URL.

## Partie V

## 

WinHTTrack est un aspirateur de sites web facile d'utilisation et [libre](http://www.gnu.org/philosophy/free-sw.fr.html) ([GPL](http://www.linux-france.org/article/these/gpl.html), logiciel libre).

Il permet de télécharger un site web d'Internet vers en construisant récursivement tous les répertoires, récupérant html, images et fichiers du serveur vers son ordinateur. HTTrack réorganise la structure des liens en relatif. HTTrack peut aussi mettre à jour un site existant, ou continuer un téléchargement interrompu. Le robot est entièrement configurable, avec un système d'aide intégré.

Source : <https://httrack.com/page/1/fr/index.html>