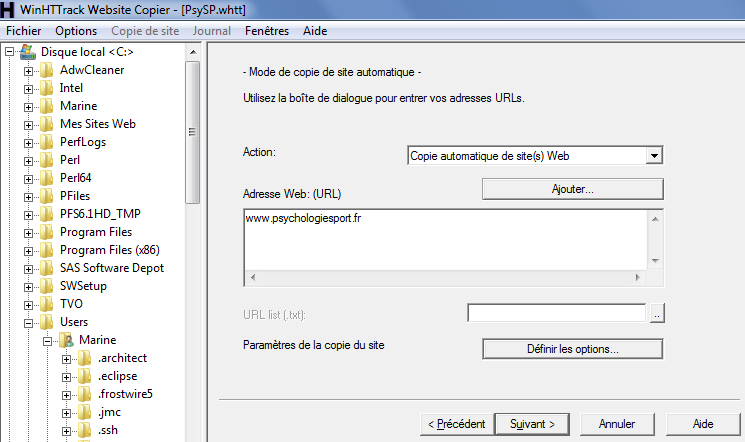
Au cours de nos recherches, on a trouvé un site internet (www.psychologiesport.fr) intéressant pour notre domaine. Ce site est composé d’une centaine d’articles concernant la psychologie dans le sport. On a remarqué en naviguant sur le site que certains articles étaient des liens vers d’autres sites,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

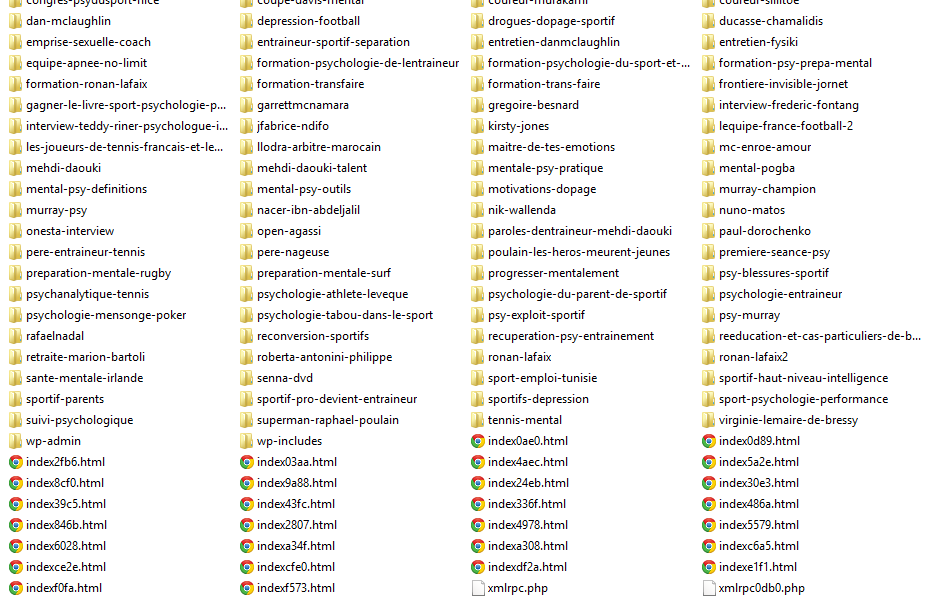


Figure 1. Capture écran d'un article de www.psychologiesport.fr dont on veut récupérer le lien vers un autre site

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



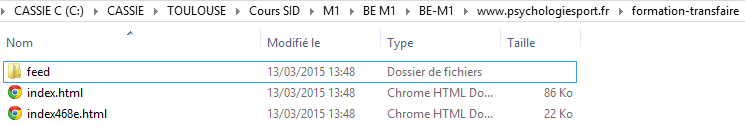
Nous avons choisi de ne récupérer que les pages du site en indiquant que la limite de profondeur externe est de 0. Ainsi HTTRACK n’est pas allé sur les sites internes autres que www.[psychologiesport.fr](http://www.psychologiesport.fr). Après avoir récupéré tous les articles du web grâce au logiciel HTTRACK, nous devions trier ces résultats pour en retirer les articles utiles.



En étudiant les résultats retournés, sous forme de dossiers (un dossier par article), nous avons remarqué que seuls les fichiers HTML nommés index.html étaient utiles.

Les fichiers en questions n’étaient pas stockés dans un dossier spécifique.

Par exemple, le fichier index.html pouvaient être présent directement à l’ouverture du dossier correspondant à l’article.



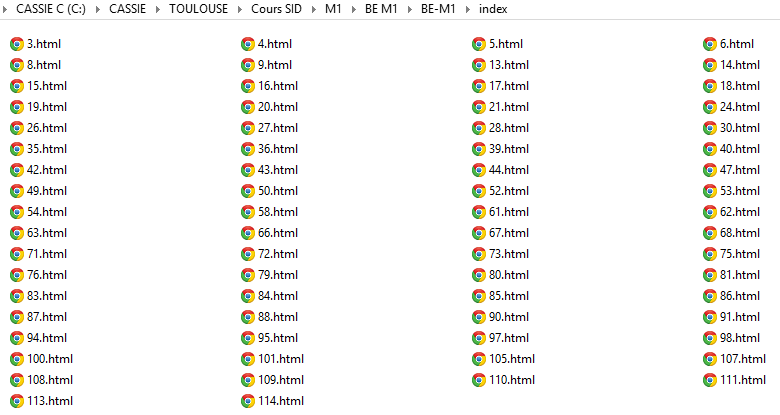
Dossier correspondant à l’article

Fichier HTML à ne pas récupérer

Fichier HTML à récupérer

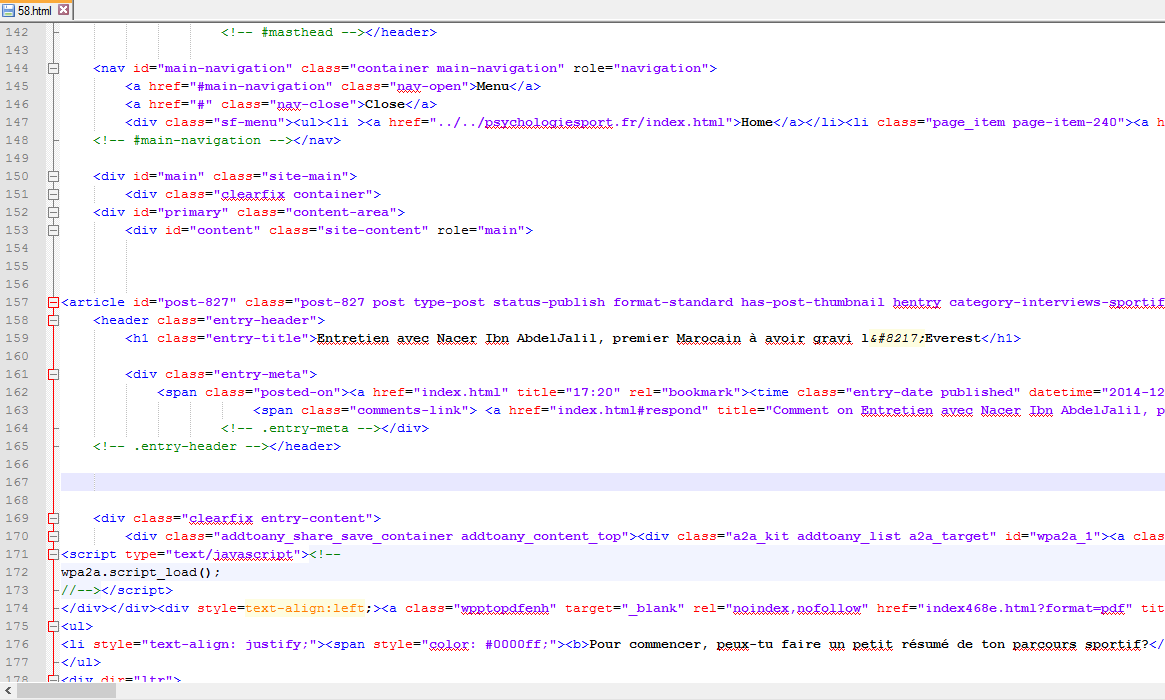
D’une autre manière cet index.html pouvait se retrouver dans un sous-dossier ou encore un sous-sous-dossier. L’emplacement était aléatoire. De plus, dans chaque dossier, un sous-dossier nommé « feed » est présent, avec, à l’intérieur un fichier « index.html » vide à ne pas garder.

Un programme perl a donc été codé : « lebon.pl ». Ce programme parcourt tous les articles (donc tous les dossiers) et récupère le bon fichier « index.html » pour le stocker dans un autre dossier qui contiendra tous les « index.html » de chaque article. Par ailleurs, chaque fichier récupéré est renommé avec un nombre incrémentant pour chaque « index.html ».



Les fichiers index.html rapatriés et renommés, une récupération du texte utile de ces pages était primordiale.

La forme de nos fichiers HTML ressemble à l’image suivante :



Nous avons réussi à identifier que le texte des articles (souligné en orange) était stocké entre les balises <article>…</article> (encadré rouge sur la figure).

Un programme « Extract.pl » a été mis à notre disposition pour extraire le texte entre certaines balises dans un fichier .html. Ce programme nous a servi de base et, par suite, l’approprier pour notre cas (fichier « ExtracModif.pl »). Pour simplifier les extractions des informations désirées et ne pas augmenter le risque de perte d’informations, nous avons réduits tous les fichiers HTML sur une seule et même ligne grâce au programme perl « ligne.pl ».

Les titres (balises <title>…</title>), contenus des articles (<article>…</article>) ainsi que les dates de publication (<time>…</time>) ont été stockés dans un fichier texte résultat. Ces résultats serviront à une future analyse sémantique des articles.

## 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>