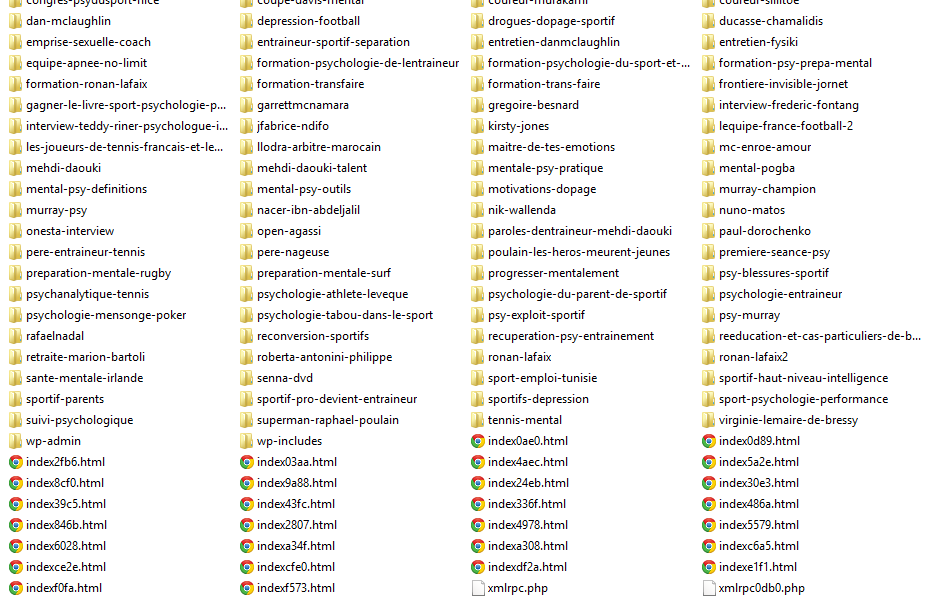
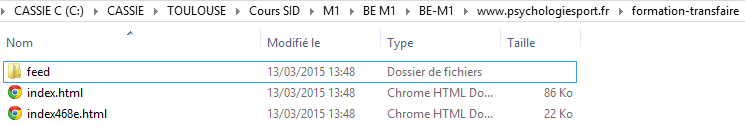
Après avoir récupéré tous les articles du web grâce au logiciel HTTRACK, nous devions trier ces résultats pour en retirer les articles utiles.



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.



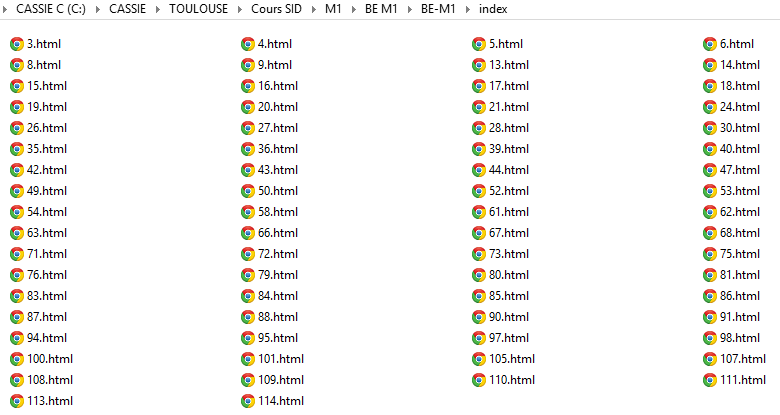
Dossier correspondant à l’article

Fichier HTML à ne pas récupérer

Fichier HTML à récupérer

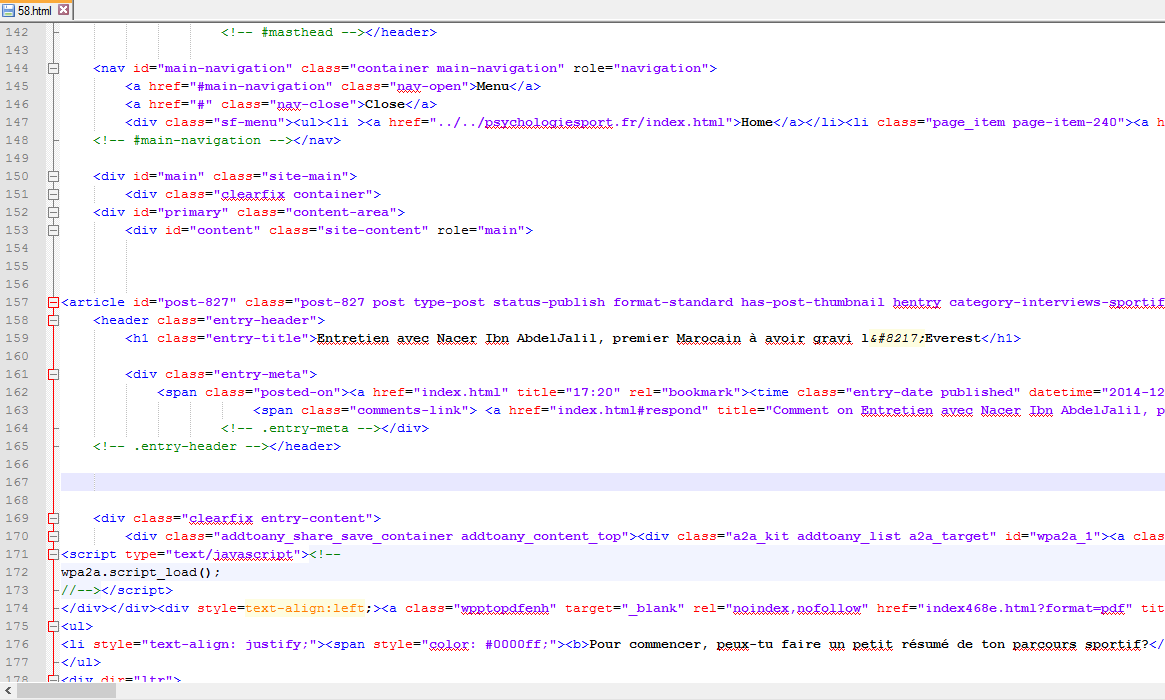
D’une autre manière cet index.html pouvait se retrouver dans un sous-dossier ou encore un sous-sous-dossier. L’emplacement était aléatoire. De plus, dans chaque dossier, un sous-dossier nommé « feed » est présent, avec, à l’intérieur un fichier « index.html » vide à ne pas garder.

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.

FUSIONS DES BASES DE DONNEES

Le sujet de notre analyse empruntant plusieurs axes thématiques tels que le sport ou encore la psychologie, nous a permis de récupérer des articles répartis sur plusieurs bases de données, plus précisément depuis trois bases :

* Pubmed : spécialisée dans la santé
* PubPsych : spécialisée dans la psychologie
* ProQuest : spécialisé dans la psychologie

Or, les formats des données récupérées différaient suivant la base de données.

|  |  |  |
| --- | --- | --- |
| Pubmed | PubPsych | ProQuest |
| IDENTIFICATEURS |  |  |

Notre objectif était donc de :

* convenir d’un paramétrage commun aux trois bases de données
* écrire un programme perl (« bd.pl ») pour récupérer les identificateurs (certains devaient être renommés) et leurs informations associées
* Fusionner les trois bases de données en une commune pour faciliter l’intégration sur le logiciel « Tétralogie »

Lors de la fusion des trois bases nous voulions tout de même pourvoir différencier les articles venant des différentes bases. Pour ce faire, nous avons convenu d’adopter un code spécifique à chaque base identifiant chaque article.

Les numéros articles provenant de la base PubMed ne changent pas tandis que ceux de PubPsych et ProQuest ont respectivement les caractères ‘PP’ et ‘PQ’ ajoutés.

|  |  |  |
| --- | --- | --- |
| PubMed | PubPsych | ProQuest |
| PMID- 25648198  PMID- 25631930  PMID- 25617538 | PMID- PP91  PMID- PP92  PMID- PP93 | PMID- PQ201  PMID- PQ202  PMID- PQ203 |

Ainsi, grâce à cette convention nous pouvons savoir de quelle base de données provient l’article en question.

Ci-dessous, les conventions adoptées pour les trois bases de données :

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | PubMed | PubPsych | ProQuest | Après renommage |
| Titre | TI | TI | Titre | TI |
| Auteur | AU | AU | Auteur | AU |
| Date de publication | DP | PY | Date de publication | DA |
| Langue | LA | LA | Langue | LA |
| Abstract/Résumé | AB | ABHR | Résumé | AB |
| Mots clés | OT+MH | CTEP | Sujet + Identificateur | MC |
| Adresse | AD | CS | Auteur | AD |
| Pays |  | COU |  | PY |
| Email |  | EMAILO | Adresse courriel | EM |