1. **Création du fichier *basededonnees***

Une fois la base fusionnée, nous avons créé le fichier *basededonnees*. Nous nous sommes inspirés d’un fichier créé lors des travaux pratiques de datamining. Pour chaque variable, nous avons cherché les séparateurs entre les termes. Ensuite, nous avons cherché toutes les balises présentes dans la base de données complète, pour les insérer dans ce fichier afin qu’elles ne soient pas utilisées comme des termes appartenant à une balise qui nous intéresse lors de l’exploitation.

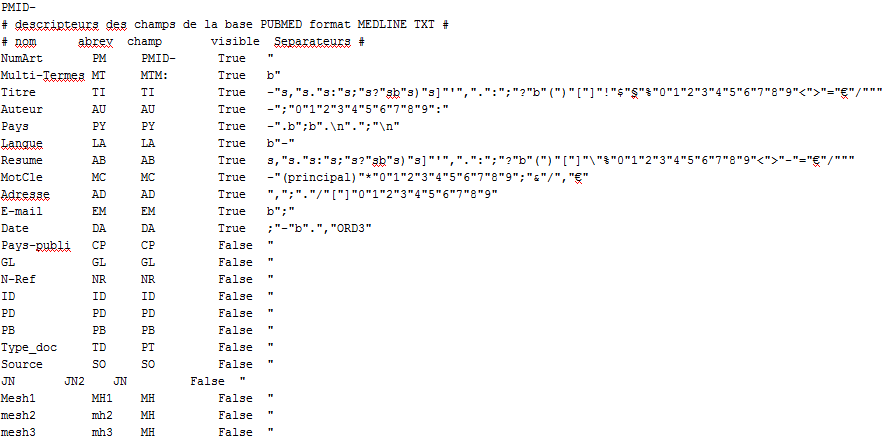


Figure  : Capture d’écran du fichier *basededonnees*.

*Difficultés* : Les trois bases utilisées n’utilisent pas forcément la même mise en forme. Par exemple, certaines bases, pour un article, voient les auteurs confinés dans une seule balise auteur (AU), et d’autres dans plusieurs balises auteur (AU).

1. **Création du fichier de mots vides**

Pour parfaire l’analyse, nous avons créé le fichier dit de mots vides. C’est un fichier recensant les mots qui n’ont pas d’utilité pour l’analyse. Comme, nous avons des articles qui sont écrits essentiellement en trois langues (français, anglais et allemand), nous avons rempli ce fichier avec des mots vides de ces trois langues. Nous avons appelé ce fichier *VIDE.Filtre*.



Figure : Extrait du fichier de mots vides en trois langues *VIDE.Filtre*

1. **Fréquences absolues**

Pour chaque variable, nous avons effectué un comptage à l’aide du logiciel Tetralogie. Pour chaque comptage, trois fichiers sont créés :

* Un fichier dont l’extension est *.ind* : recense tous les termes différents présents sous la balise.
* Un fichier dont l’extension est *.indA* : recense tous les différents termes par ordre alphabétique ainsi que la fréquence d’apparition.
* Un fichier dont l’extension est *.indF* : recense tous les différents termes ainsi que leur fréquence d’apparition et classés par ordre décroissant d’apparition.

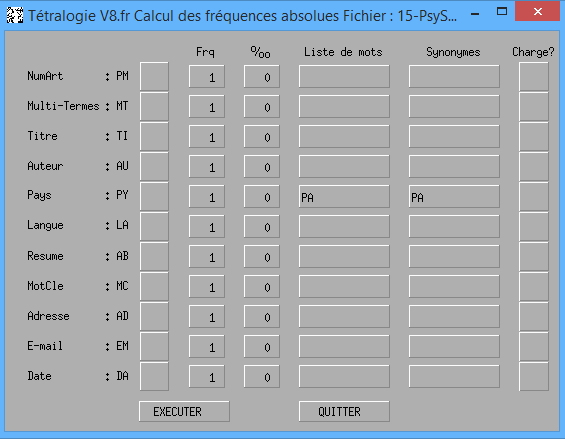


Figure : Capture d'écran de la fenêtre d'exécution des fréquences absolues

* 1. **Champs titre, résumé et mots-clé**

Nous avons réalisé le comptage en retirant les mots vides pour les champs titre et résumé, ainsi qu’un comptage simple sur le champ mots-clé.

* 1. **Champs langue et pays**

Nous avons créé un fichier *LA.Filtre* qui retire certains termes parasites (ici, certaines adresses URL) et un fichier de synonymes *LA.Syn* qui permet de regrouper les synonymes pour les comptages.



Figure : Fichier de synonymes *LA.Syn*

Pour le champ pays, nous avons utilisé un fichier filtre déjà créé *PA.Filtre* qui contient la liste des pays, puis nous avons complété le fichier *PA.Syn* avec les synonymes que nous avons rencontrés dans notre base.

* 1. **Champs auteur**

Pour le champ auteur, nous avons extrait ce qui était présent sous les balises AU. Ensuite, nous avons crée un fichier *auteur.Filtre* dans lequel nous avons inséré tous les auteurs rencontrés. Nous avons ensuite appliqué ce filtre positivement pour récupérer uniquement les auteurs.

Enfin, nous avons établi le fichier de synonymes *AU.Syn* que nous avons corrigé par la suite.

*Difficulté*: Rassembler les auteurs qui sont placés de manière différente selon les bases.

* 1. **Champ date**

Pour le champ date, nous avons créé un fichier *DA.Filtre* contenant toutes les années de 1900 à 2015. Nous avons ensuite appliqué ce filtre positivement. Ensuite, nous avons créé un fichier de synonymes *DA.Syn* associant chaque année à une période.

Difficulté : Extraire les années des différents formats de date rencontrés.

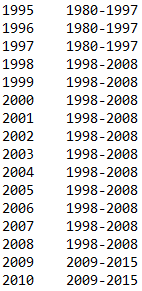


Figure : Fichier *DA.Syn* associant une date à une période.