**Mode opératoire de la structuration des données graph (Gephi)**

**R. Sauvayre**

Le principe est de rassembler et organiser des données pour que je puisse en construire un graph avec le logiciel Gephi. J’étudie les liens existants entre des auteurs publiant dans des articles scientifiques. Je m’intéresse 1/ aux liens de coécriture : personnes qui co-écrivent un ou plusieurs articles ensemble et 2/ aux liens de citations : des personnes qui citent d’autres personnes dans leur bibliographie.

Il s’agit donc de structurer ces liens de sorte qu’ils soient lisibles par Gephi et ce au format .csv. Ce logiciel demande deux fichiers : un fichier « Edge » contenant les liens entre les données et un fichier « Nodes » contenant les intitulés ou données supplémentaires.

Vous trouverez ci-joint plusieurs fichiers :

1/ le fichier de données « Complete Data » en .xls

2/ un fichier exemple « Nodes » pour que vous puissiez voir comment structurer les liens

3/ un fichier exemple « Nodes » pour que vous puissiez voir comment structurer les liens et données

Les liens entre coauteurs portent le nom « CoALink » et doit être inscrit dans la colonne « Label » d’un fichier Edges.csv

Les liens entre auteurs citant ou lien de citation portent le nom« CitLink »et doit être inscrit dans la colonne « Label » d’unfichier Edges.csv

**Le fichier Edge**

Une ligne du tableur correspond à un lien. Il y a deux types de liens possibles. Voici comment remplir le fichier Edge :

1. Lien de coauteur : L’ID de l’auteur 1 qui a coécrit avec l’auteur 2 est inscrit dans la colonne « Source » et l’ID de l’auteur 2 est inscrit dans la colonne « Target » ; le « Type » est « **Undirected** » ; le poids « Weight » est toujours 1 et le « Label » est « **CoALink** ». Cf. plus bas pour trouver les données.
2. Lien de citation : L’ID de l’auteur 1 qui cite l’auteur 2 est inscrit dans la colonne « Source » et l’ID de l’auteur 2 est inscrit dans la colonne « Target » ; le « Type » est « **Directed** » ; le poids « Weight » est toujours 1 et le « Label » est « **CitLink** ». Cf. plus bas pour trouver les données. Il faut veiller à ce que l’auteur citant soit dans la colonne Source et l’auteur cité soit dans la colonne Target, car cette fois-ci les liens sont dirigés (Directed) et que cela aura une incidence dans le traitement des données.
3. enfin, comme dans l'exemple ci-joint, il y a des **autocitations** (le même ID d'auteur se trouve dans « source » et dans« target »). Il faut **supprimer** ce lien ou empêcher qu'il soit créé.

**Le fichier Nodes**contient alors :

* L’identifiant des auteurs (ID author) à rechercher dans le fichier Complete\_Data, colonne « ID author »(colonne K)
* Le nom des auteurs (Label) à rechercher dans le fichier Complete\_Data, colonne « Nom auteur : complet 2 (sans accent) »(colonne N)
* L’identifiant de l’article dans lequel ils ont écrit (ID Paper)est à rechercher dans le fichier Complete\_Data, colonne « ID Paper »(colonne A)

Comment trouver ces liens dans le fichier Complete\_Data ?

* Le lien de coauteur : lorsque les identifiants de la colonne « ID Paper » (colonne A) sont identiques, cela signifie que les auteurs ont coécrit ensemble. Il faut alors lier chaque coauteur comme indiqué dans le point 1. L’ID de chaque auteur est à trouver dans la colonne « ID author » (colonne K) ;
* Le lien de co-citation : l’auteur citant est à trouver dans la colonne « ID author » (colonne K) et les auteurs cités sont à trouver dans les colonnes « Cited\_author » (colonne BT à CS).

Un autre point utile : on peut vérifier que les liens sont bien structurés lorsque l'on charge les deux fichiers dans Gephi. Il indique systématiquement le nombre de liens trouvés (liens dirigés et non dirigés).