Intervention de Baptiste Coulmont

Le corpus aujourd'hui présenté ne porte pas uniquement sur des corpus de textes. La présentation est fondée sur un lancement de R en direct. La présentation se fait sous la forme de lignes de codes. Un effort premier est nécessaire pour comprendre le codage sous R. Parfois le codage est assez facile, parfois c'est beaucoup plus long. L'intérêt de R est la sauvegarde sous format txt.

Le premier exemple porte sur les noms des lieux, les toponymes. La *library maptools* permet de faire des cartes en .shp (*shapefile*). Le fichier GEOFLA de l'IGN est chargé. Les noms des communes en France présente certaines caractéristiques en termes de toponymes : saint, sur, sous... Ici la question des lettres finales (du y, du c et du m) est posée pour associer des toponymes à des régions privilégiées. Le m final se trouve en Alsace mais aussi dans le Pas-de-Calais. Les communes avec un c final se trouvent dans le sud-ouest et en Bretagne. Pour le y final, la géographie est moins visible. La carte produite sous R est exportable sous différents formats, notamment pdf.
Le deuxième exemple porte sur un fond de carte du monde pour représenter les voyages d'un pasteur afro-américain dans le monde. Des connexions sont clairement visibles : l'Amérique du Nord, l'Europe et une partie de l'Afrique où l'évangélisme est en plein essor. Cela repose sur des déclarations sur un site Internet recodés sur une table .csv avec le toponyme, la latitude et la longitude.
Open Street Map est un équivalent de Google map créé par des utilisateurs. Il est alors possible de télécharger des .shp. Une étude de cas sur l'Ile-de-France est présentée. En 2007, à Paris, une loi interdit l'installation des *sex shops* à moins de 200 mètres d'un établissement scolaire. Il s'agit de spatialiser les lieux qui pourraient alors en accueillir. Tous les calculs de type *buffer zones* d'interdiction de 200 mètres sur les 930 établissements scolaires sont faits sous R. Cet amendement restreint de fait l'installation de *sex shops*. Les connaissances en termes de SIG pour les entrepreneurs de *sex shops* sont parfois faibles, et un bureau d'études coûteux.
Un fichier contient les chiens domestiques français (11 millions) avec leur race et leur nom. Les données sont représentées sous la forme d'un réseau via la *library* igraph. Selon la race de chiens, les gens ont tendance de donner le même prénom. Il n'y a pas seulement des races des chiens, mais les propriétaires choisissent souvent des prénoms ethniques. La visualisation des données permet de faire passer des idées facilement.

Il est très facile sous R d'aspirer des données twitter pour avoir des données utilisables, notamment quand sur twitter il y a la localisation de la question.
L'interface Rstudio fonctionne sous tous les OS (windows, linux, mac). Le langage R permet de multiplier les compétences.