A. Problématique

Dans de nombreux cas de figure, nous ne disposons pas de toutes les données suffisantes pour une analyse dans une seule source de données. Il est alors nécessaire de procéder à ce que nous appelons la mise en relation des données. Pour cela, nous nous appuyons sur des clés, à savoir des informations partagées entre les différentes sources de données à relier entre elles, de façon à ce que l'ensemble des sources forme un tout homogène, et surtout exploitable.

Nous allons étudier ci-dessous les différents moyens pour créer une relation entre plusieurs sources de données, ainsi que montrer les pièges d'un manque de relation entre les données, ce qui peut amener à des calculs faux.

B. Mise en place de relation

1. Présentation des données

Nous allons utiliser deux sources de données en provenance de l'INSEE pour illustrer les concepts de clé et de relation, et voir un exemple d'application. La première correspond aux données géographiques des communes de France, tandis que la seconde reprend les résultats des élections présidentielles 2007.

La première source donne, pour toutes les communes de France, des informations sur le département et la région d'appartenance, ainsi que la population de la commune. Son format est celui d'un fichier CSV (*Comma Separated Values*), bref un simple fichier texte dans lequel les données sont séparées traditionnellement par des virgules, comme l'indique l'acronyme, bien que dans notre cas précis le séparateur soit le point-virgule.

Voici les dix premières lignes de cette source de données :

```
#"Département commune";"Libellé de
commune";"Région";"Département";"Arrondissement";"Canton ville";"Zone
d'emploi";"Unité urbaine";"Aire urbaine";"Espace urbain";"Tranche d'aire
urbaine";"Tranche de commune";"Taille des unités urbaines";"Tranche
détaillée d'unité urbaine";"Bassin de vie";"Etablissement public à
fiscalité propre";"Nature d'établissement public";"Catégorie de
ZAUER";"Population sans doubles comptes 1999";"Population municipale
2007"
```

```
#"CODGEO";"LIBGEO";"REG";"DEP";"ARR";"CV";"ZE1990";"UU1999";"AU1999";"EU1
999";"TAU1999";"TC1999";"TUU1999";"TDUU1999";"BV";"N° EPCI";"Nature
EPCI";"ZAUER1999";"PSDC99";"Pop_mun_2007"

"01001";"L'Abergement-
Clémenciat";"82";"01";"012";"0110";"8203";"01000";"999";"1C";"00";"1";"0"
;"05";"01093";"240100644";"CC";"3";728;804

"01002";"L'Abergement-de-
Varey";"82";"01";"011";"0101";"8210";"01000";"999";"1C";"00";"0";"0";"03"
;"01004";"240100883";"CC";"3";168;195

"01004";"Ambérieu-en-
Bugey";"82";"01";"011";"0101";"8210";"01303";"327";"1C";"01";"3";"3";"31"
;"01004";"240100883";"CC";"1";11436;12696

"01005";"Ambérieux-en-
Dombes";"82";"01";"012";"0130";"8203";"01000";"002";"1C";"09";"1";"0";"06
";"69264";"240100735";"CC";"2";1408;1544

"01006";"Ambléon";"82";"01";"011";"0104";"8209";"01000";"307";"1C";"02";"
0";"0";"02";"01034";"240100354";"CC";"2";86;125

"01007";"Ambronay";"82";"01";"011";"0101";"8210";"01000";"999";"1C";"00";
"1";"0";"07";"01004";"240100883";"CC";"3";2146;2263

"01008";"Ambutrix";"82";"01";"011";"0117";"8210";"01000";"999";"1C";"00";
"1";"0";"05";"01004";"240100883";"CC";"3";586;649

"01009";"Andert-et-
Condon";"82";"01";"011";"0104";"8209";"01000";"307";"1C";"02";"0";"0";"04
";"01034";"240100354";"CC";"2";275;325
```

La seconde source présente les résultats des deux tours de l'élection présidentielle de 2007. Nous ne nous intéressons pour cet exemple qu'au deuxième tour. Cette source de données est également présentée avec des informations par commune, mais avec une décomposition supplémentaire, à savoir les différents bureaux de vote, qui peuvent être multiples pour une commune.

Pour fixer les idées, voici également un extrait de cette source, elle aussi au format CSV :

```
--  Résultats par bureau de vote Présidentielles 2007 Tour 2
-
-
-- Champ 1 : N° tour
-- Champ 2 : Code département
-- Champ 3 : Code de la commune
-- Champ 4 : Nom de la commune
-- Champ 5 : N° de bureau de vote
-- Champ 6 : Inscrits
-- Champ 7 : Votants
-- Champ 8 : Exprimés
-- Champ 9 : N° de dépôt du candidat
-- Champ 10 : Nom du candidat
-- Champ 11 : Prénom du candidat
-- Champ 12 : Code nuance du candidat
-- Champ 13 : Nombre de voix du candidat
--
2;01;001;L'Abergement-
Clémenciat;0001;596;534;512;8;ROYAL;Ségolène;ROYA;197
2;01;001;L'Abergement-
Clémenciat;0001;596;534;512;12;SARKOZY;Nicolas;SARK;315
2;01;002;L'Abergement-de-Varey;0001;205;183;178;8;ROYAL;Ségolène;ROYA;76
2;01;002;L'Abergement-de-
Varey;0001;205;183;178;12;SARKOZY;Nicolas;SARK;102
2;01;004;Ambérieu-en-Bugey;0001;1077;853;810;8;ROYAL;Ségolène;ROYA;394
2;01;004;Ambérieu-en-Bugey;0001;1077;853;810;12;SARKOZY;Nicolas;SARK;416
2;01;004;Ambérieu-en-Bugey;0002;912;726;699;8;ROYAL;Ségolène;ROYA;320
2;01;004;Ambérieu-en-Bugey;0002;912;726;699;12;SARKOZY;Nicolas;SARK;379
```

2. Formatage initial

Vous pouvez remarquer que toutes les données du premier fichier soient entourées de guillemets. C'est une pratique qui se retrouve souvent dans les fichiers CSV contenant potentiellement des apostrophes, mais en général réservée aux données textuelles. Nous n'en aurons pas besoin ici. De même, les symboles # préfixant les premières lignes servent à indiquer que la ligne ne contient pas une donnée, mais une remarque sur le contenu, qui ne doit pas être prise en compte. Le module d'import de PowerPivot pouvant être paramétré pour supporter une ligne de titre, nous modifierons également cette partie du fichier.

Quant à la seconde source de données, il faudra supprimer les dix-sept premières lignes correspondant à de la documentation des colonnes, mais il peut être intéressant de réécrire ceci sous forme d'une ligne additionnelle de données.

Les fichiers déjà formatés sont livrés dans les téléchargements associés au livre, dans le sous-répertoire INSEE. Vous pouvez donc vous passer de cette section technique si vous le souhaitez.

Les manipulations à réaliser pour le premier fichier sont les suivantes :

- Ouvrez le fichier **Appartenance-géographique-des-communes-01-01-2009.csv** dans un éditeur de texte. Vu la taille du fichier, il est recommandé d'utiliser un éditeur plus moderne que notepad (inclus dans Windows), par exemple notepad++ (http://notepad-plus-plus.org).
- Utilisez la fonction **Rechercher / Remplacer** pour localiser les caractères guillemets doubles (") et les remplacer par une chaîne vide.
- Exécuter l'opération **Remplacer tout**. Cette opération peut prendre jusqu'à quelques minutes, vu le nombre d'occurrences.
- Supprimez la seconde ligne du fichier.
- Supprimez le symbole dièse (#) de la première ligne du fichier.
- Sauvegardez.

En ce qui concerne le second fichier, la marche à suivre est la suivante :

- Supprimez les trois premières lignes.
- Supprimez la ligne entre les titres et les données.

- Juste avant la première ligne de données, fusionnez les titres exprimés sur plusieurs lignes en une seule. Plutôt que de la saisir, vous pouvez supprimer simplement les caractères et saut de lignes superflus. Attention toutefois à bien utiliser des points-virgules pour séparer les différents titres.
- Au final, vous devez obtenir un fichier dont le début est comme suit :

```
N° tour;Code département;Code de la commune;Nom de la commune;N° de
bureau de vote;Inscrits;Votants;Exprimés;N° de dépôt du candidat;Nom du
candidat;Prénom du candidat;Code nuance du candidat;Nombre de voix du
candidat
2;01;001;L'Abergement-
Clémenciat;0001;596;534;512;8;ROYAL;Ségolène;ROYA;197
2;01;001;L'Abergement-
Clémenciat;0001;596;534;512;12;SARKOZY;Nicolas;SARK;315
2;01;002;L'Abergement-de-Varey;0001;205;183;178;8;ROYAL;Ségolène;ROYA;76
2;01;002;L'Abergement-de-
Varey;0001;205;183;178;12;SARKOZY;Nicolas;SARK;102
2;01;004;Ambérieu-en-Bugey;0001;1077;853;810;8;ROYAL;Ségolène;ROYA;394
2;01;004;Ambérieu-en-Bugey;0001;1077;853;810;12;SARKOZY;Nicolas;SARK;416
2;01;004;Ambérieu-en-Bugey;0002;912;726;699;8;ROYAL;Ségolène;ROYA;320
2;01;004;Ambérieu-en-Bugey;0002;912;726;699;12;SARKOZY;Nicolas;SARK;379
2;01;004;Ambérieu-en-Bugey;0003;954;773;733;8;ROYAL;Ségolène;ROYA;335
```

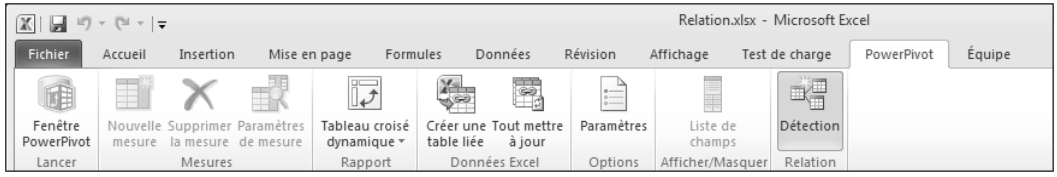

Ces manipulations de données peuvent paraître fastidieuses, mais elles sont nécessaires pour que PowerPivot les lise correctement. Le format CSV étant basé sur du texte, il est difficile d'empêcher les producteurs d'ajouter des lignes multiples de titre ou des symboles particuliers, bref de ne pas respecter strictement le format CSV. C'est tout l'intérêt d'une approche plus formalisée pour la diffusion de données, en utilisant par exemple des formats XML ou JSON, qui peuvent être validés par des grammaires plus strictes, réduisant ainsi à néant cette étape de formatage.

3. Intégration des données dans PowerPivot

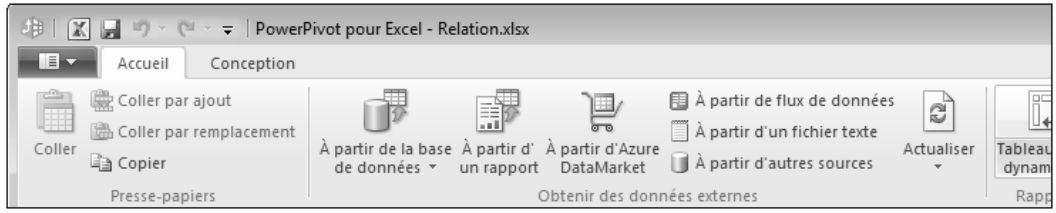
La préparation des données étant terminée, nous commençons par intégrer ces sources de données comme nous l'avons fait précédemment, à la différence que nous dupliquons l'étape d'import.

La marche à suivre est la suivante :

- Lancez Excel.
- Dans le ruban, sélectionnez l'onglet **PowerPivot**.



- Lancez la **Fenêtre PowerPivot** par l'icône associée.
- Utilisez la commande **À partir d'un fichier texte** du groupe **Obtenir des données externes**.



- Dans l'assistant qui apparaît, utilisez le bouton **Parcourir** et sélectionnez le premier fichier **Appartenance-géographique-des-communes-01-01-2009.csv**. Attention, vous aurez peut-être à modifier le type de fichiers affichés (en bas à droite de la boîte de dialogue) pour voir apparaître les **fichiers CSV (*.csv)**.
- Cliquez sur **Ouvrir**.