Université de Technologie de Compiègne

IC05

Rapport de projet

Medifiles

Automne 2016

Quentin GRAS - William BIRGEL - Kyâne PICHOU

12 janvier 2017

Table des matières

1 Introduction

Dans le cadre de l'article L.1453-1 du code de la santé publique, les entreprises produisant ou commercialisant des produits médicaux ou assurant des prestations associées à ces produits sont tenues de rendre publique l'existence des conventions qu'elles concluent avec différents acteurs de la santé (médecins, étudiants, associations, etc.).

De ce fait, les avantages perçus par des acteurs du domaine de la santé sont accessibles sur le site https://www.transparence.sante.gouv.fr/. Ce site permet de faire des recherches à l'aide d'un formulaire, suivant un certain nombre de critères. Cependant ceci ne permet pas de répondre facilement à certaines questions (qui touche le plus ? qui donne le plus ? quel secteur est le plus avantagé ?) et ne permet pas d'avoir une vision d'ensemble des données en question. En 2015, l'association Regards Citoyen a analysé une partie de ces données pour en sortir des statistiques génériques. Depuis, et dans le cadre de la mission Etalab, les données sont proposées sous la forme de fichiers CSV.

L'objectif de notre projet est donc de récupérer et de traiter ces données, pour voir ce que l'on peut obtenir comme informations. Idéalement, nous aimerions développer des scripts permettant de faire une analyse automatisée des données, dans l'éventualité d'une poursuite ou d'une passation du projet.

2 Traitement des données

2.1 Préambule

Pour commencer nous avons récupéré les fichiers CSV disponibles sur le site DataGouv https://www.data.gouv.fr/fr/datasets/transparence-sante-1/. Les données sont fournit sous la forme de 3 fichiers CSV ainsi que quelques fichiers de présentation rapide du format des données.

Un premier fichier contient la liste des entreprises présentes dans la base de données des déclarations. Un second fichier contient la liste des avantages (informations sur les bénéficiaires, l'entreprise et les montants), et le dernier fichier contient la liste des conventions passés entre une entreprise et un bénéficiaire. Les informations actuellement disponibles pour les conventions sont partielles (pas de montant par exemple). Ceci devrait possiblement changer en Avril 2017, mais c'est pour cela que nous avons porté notre attention uniquement sur les avantages (ou cadeaux).

Notre objectif sera d'extraire les données des fichiers CSV pour représenter l'ensemble sous la forme d'un graphe. Ce graphe sera chargé dans une base de donnée, orientée graphe, Neo4j. Le schéma sera très simple : les entreprises et bénéficiaires seront des noeuds et chaque avantage sera un lien (entreprise vers bénéficiaire). Chaque objet (noeuds et liens) aura des attributs correspondant aux informations fournit dans les fichiers CSV (nom, adresse, montant, etc.).

2.2 Étude du format des données

2.2.1 Entreprises

Le CSV contenant les entreprises est relativement bien présenté. Pour chaque entreprise nous avons un identifiant, un nom, une adresse, et un secteur d'activité. Après nettoyage des quelques caractères mal encodé dans le nom, on constate que les informations sont assez bien formatées :

— l'identifiant d'entreprise est bien unique

— le secteur d'activité est désigné par un code. Ce n'est pas un champ libre (ceci permet donc de faire des tri par secteur).

2.2.2 Avantages

Détermination des informations intéressantes

Le fichier contenant les avantages est très volumineux (plus de 6 millions de lignes) et la façon de formater les données est un peu moins standard. Chaque ligne représente un avantage, mais les attributs ne sont pas systématiquement renseignés. En fonction du statut de la personne physique ou morale, les différents champs peuvent donc être vides ou remplis différemment.

Un premier travail a été de déterminer quelles étaient les informations dans chaque champs et leur format. Parmi les 36 attributs de chaque avantage, tous ne nous intéressent pas pour la suite de l'étude. On compte ainsi dans ce qui nous intéresse le plus : nom, prénom et adresse du bénéficiaire, identifiant de l'entreprise associée, montant, date et nature de l'avantage.

Analyse du format

L'analyse des données nous a rapidement fait constater que les champs libres (code postal, nature de l'avantage) ne sont pas exploitables de manière automatique. Par exemple, certains code postaux ne sont pas renseignés et contiennent parfois tout simplement du texte. De même le champ concernant la nature de l'avantage est clairement du texte libre, ce qui ne permet pas de regrouper facilement les avantages en différents types.

D'autre part, les bénéficiaires n'ont malheureusement pas d'identifiant unique. De plus, un bénéficiaire pouvant avoir plusieurs spécialités, établissements ou adresses, il n'est pas possible d'identifier chaque personne de manière certaine. Compte tenu du volume de donnée, nous avons fait le choix d'identifier les personnes à partir de la combinaison nom et prénom, le nombre d'erreurs de collision étant potentiellement faible.

2.3 Création d'une base de graphe

2.3.1 Format de la base

L'objectif est de pouvoir charger les données dans une base orientée graphe, permettant de manipuler les données comme étant un graphe.

Les entreprises seront notre premier type de noeud (type *Entreprise*). Celui-ci a les attributs suivants :

identifiant : un identifiant unique.

nom : la dénomination sociale de l'entreprise.

code_pays : un code pays à deux lettres (FR, US, etc.).

code_secteur : un code secteur. La correspondance code<->libellé du secteur est disponible dans la documentation fournit avec les fichiers CSV.

adresse : l'adresse annoncée de l'entreprise.

code_postal : son code postal.

ville : la ville correspondante.

Les bénéficiaires seront notre second type de noeud. Pour faciliter notre étude, et comme ils représentent 99.52% des bénéficiaires, nous allons uniquement traiter les cadeaux fait à des professionnels de santé et/ou des étudiants. Ceux-ci seront représentés par des noeuds de type *Personne* ayant les attributs suivants :

nom : le nom du bénéficiaire.

prénom : son prénom.

id_benef : dans certains rares cas, un identifiant pour les professionnel est renseigné.

id_type_code : lorsque qu'un identifiant est renseigné, son type est spécifié (id interne à l'entreprise, id de l'ordre des médecin, siret, etc.)

type : le type de bénéficiaire. Ici uniquement ETU (étudiant) et PRS (professionnel de santé).

Pour finir les liens auront tout les attributs d'un avantage, car ceux-ci sont généralement des champs libres (à l'exception du montant et de la date). Cependant nous avons effectué un rapide nettoyage des entrées ayant des attributs incomplets ou incompréhensibles.

2.3.2 Nettoyage des données

Avant de passer au chargement des données dans la base à proprement parler, nous avons effectué un rapide nettoyage de certains champs.

Par exemple, les nom et prénom de personnes doivent systématiquement être passés en minuscule, sans accent ni caractères spéciaux. En effet de trop nombreuses personnes ont été saisies avec une syntaxe différente d'une entreprise à une autre. On simplifie donc au maximum le texte saisi pour minimiser le risque d'erreurs.
En règle générale, dans un volume aussi important de données en saisie libre, il est préférable de simplifier au maximum le format des données (retrait de la ponctuation, des caractères spéciaux type apostrophe, etc.) pour éviter des problèmes d'encodage ou des erreurs dans nos requêtes à la base de données.

De plus, il convient de vérifier le format de certains attributs. Par exemple, tout les codes postaux ne correspondant pas à une suite de 5 chiffres (ou simplement un champ vide) sont mis de côtés, pour éviter de charger en base des informations qui n'ont pas le bon format.
On notera cependant qu'aucune solution n'a été mise en oeuvre pour l'attribut concernant la nature de l'avantage : ce champ en saisie libre a été chargé tel quel dans la base de donnée, mais pourra être traité à posteriori (analyse par expression régulière) pour essayer de donner un format standard à cet attribut.

2.3.3 Réalisation

Pour notre base de données, nous avons choisi d'utiliser Neo4j qui est la référence actuelle en base orientée graphes. C'est un moteur dont le langage de requête (Cypher) est simple mais assez puissant pour nos besoins. Pour notre projet nous avons déployé la base via Docker sur une instance AWS temporaire.

Pour charger les données, nous utiliserons des scripts en Python qui vont réaliser le nettoyage des données (voir partie 2.3.2) puis charger les données dans la base Neo4j, créant ainsi le graphe que l'on souhaite obtenir. L'ensemble des petits scripts utilisés sont disponibles sur le repository Git `https://gitlab.utc.fr/pichouky/medifiles`.

Pour commencer on charge le CSV des 2051 entreprises, à l'aide d'un premier script. Ensuite, un second script se charge de créer dans la base tout les bénéficiaires et les cadeaux reçus. Ce traitement est très long car il y a énormément d'avantages à traiter (environ 6 millions). Nous avons fait le choix de n'ajouter dans la base que les avantages déclarés au dessus de 100 €. Ceci représente uniquement 15% des avantages (soit 898 181), mais la création de ce graphe a tout de même pris environ 60 heures.

3 Résultats

Une fois notre base Neo4j peuplée, nous avons un grand graphe que l'on peut manipuler pour extraire des informations intéressantes.

3.1 Visualisation par graphe

Pour commencer nous voulions visualiser le graphe via le logiciel Gephi. Un script a été développé pour extraire de la base Neo4j des fichiers CSV à importer à Gephi, à partir d'une requête à la base.

Nous ne prendrons que les transactions dont les montants sont supérieurs ou égaux à 1000 € du fait du trop grand nombre de données à traiter. Chaque noeuds représente un professionnel de santé ou bien une entreprise. Chaque liens représente donc une transaction passée entre le professionnel de santé et l'entreprise. La taille des noeuds représentant les entreprises est proportionnel au montant des transactions totales fait par celle-ci. Nous obtenons un graphe orienté de 26385 noeuds et de 52726 liens (voir figure 3.1). Par rapport aux données brutes de notre graphe, nous trouvons une modularité de 0.808, un degré de 1.998 et bien sûr un diamètre de 1 car les entreprises ne sont pas liées entre elles. On remarque tout de suite que deux entreprises dominantes : *Bayer HealthCare* et *Novartis Pharma* toute deux capitalisant respectivement 8260492 € et 5145499 €. Ces deux entreprises sont également celles qui possèdent le plus grand nombre de liens à savoir 3737 et 2776. Ces données concernent uniquement les dons supérieurs à 1000 €.

La quantité de données et la sur-représentation de certains industriels ne permet pas, à première vue, d'arriver à des conclusions pertinentes. Même en cassant les noeuds principaux, on ne peut observer que quelques clusters vraiment représentatif. Dans le graphe principal, les noeuds marron et bleu ciel représentent majoritairement les domaines chirurgicaux.

Pour enrichir notre étude nous avons également décidé de calculer le montant moyen perçu par les professionnels de santé par entreprises (voir figure 3.2). On remarque que celui-ci est assez constant sauf pour quelques entreprises. Après étude, il se trouve que ces dernières n'ont pas beaucoup de professionnels de santé affiliés (1 ou 2), seul MORIAS avec 16 professionnels de santé en lien leur a offert en moyenne 20000 € de cadeaux.

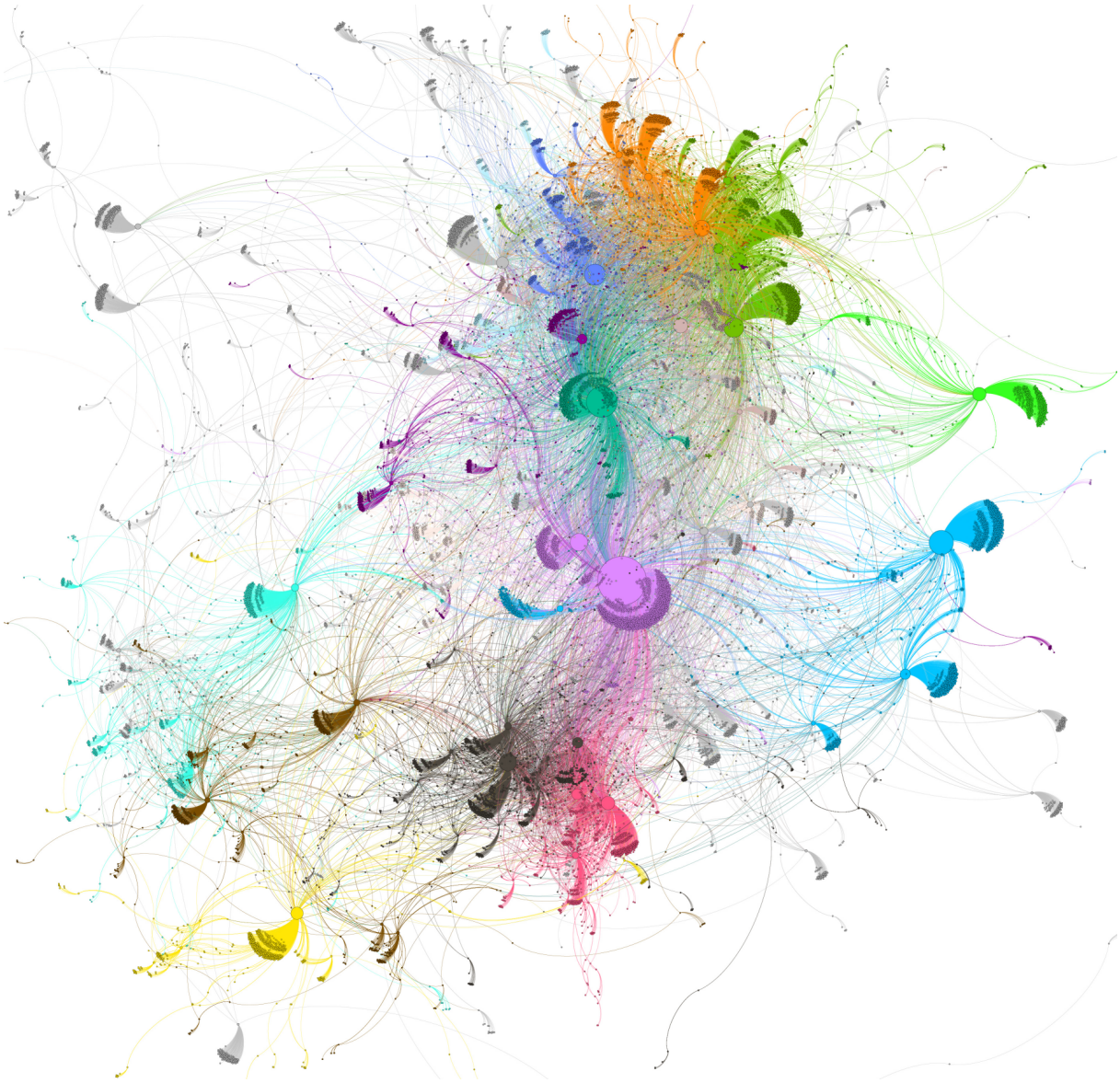

FIGURE 3.1 – Graphe des avantages

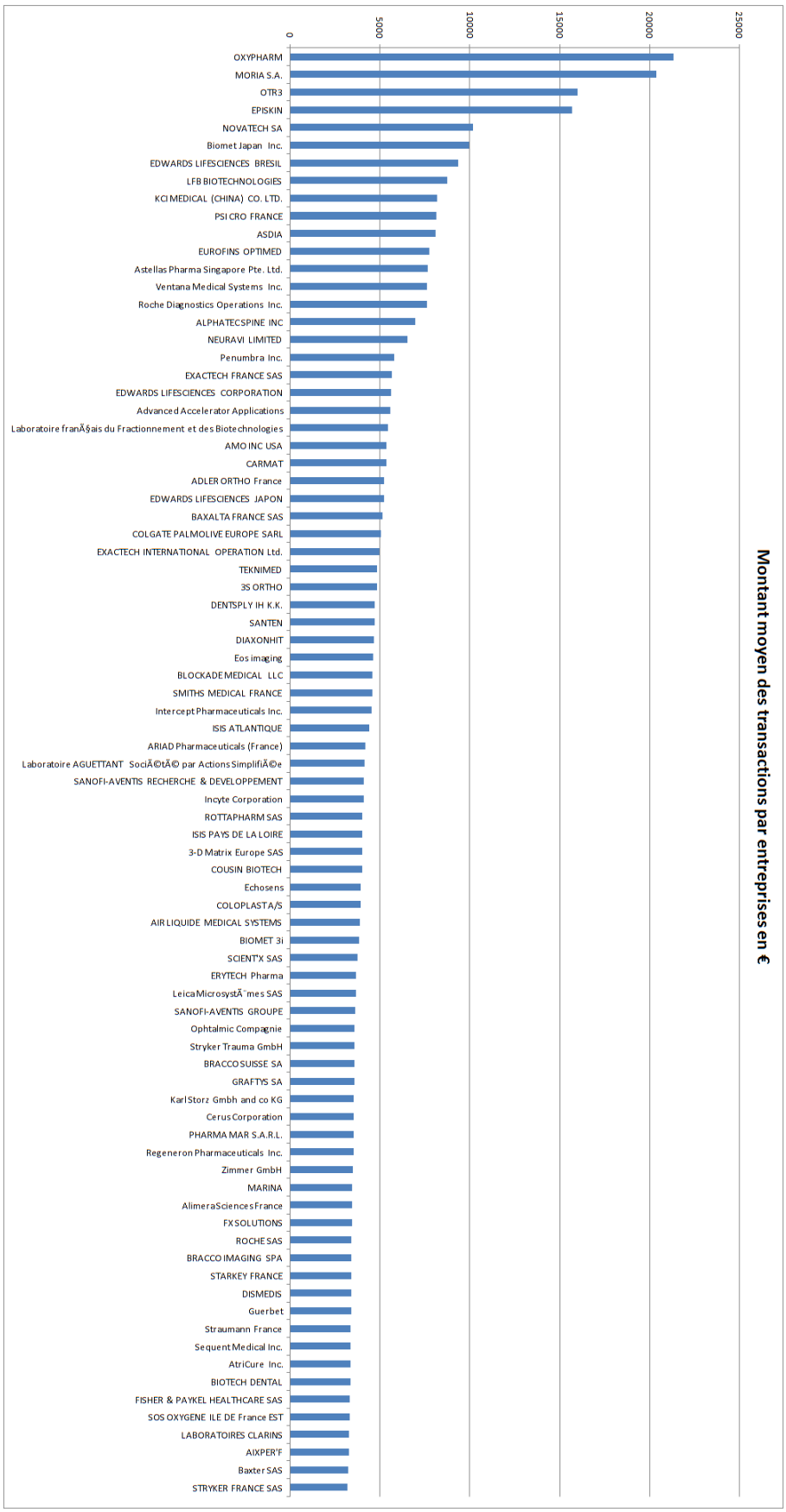

FIGURE 3.2 – Graphique des montants moyens par entreprise

3.2 Analyse globale

A partir de la base de donnée, nous pouvons sortir quelques chiffres concernant les entreprises et praticiens.

Entreprises	**Nombre de cadeaux**
NOVARTIS PHARMA SAS	44468
ROCHE SAS	29950
MSD France	27803
MEDTRONIC France S.A.S	21957
SANOFI AVENTIS FRANCE	20479
JANSSEN-CILAG	19729
Lilly France SAS	19632
LABORATOIRE GLAXOSMITHKLINE	19217
AbbVie	19068
BRISTOL-MYERS SQUIBB	18624
LES LABORATOIRES SERVIER	16562
AMGEN SAS	16107
ETHICON	15701
ASTELLAS PHARMA	14958
MERCK SERONO	14436

Table 3.1 – Entreprises faisant le plus de cadeaux

Entreprises	**Montant total (€)**
NOVARTIS PHARMA SAS	19468227
Bayer HealthCare SAS	12604822
MSD France	11907117
ROCHE SAS	11709234
MEDTRONIC France S.A.S	9326699
JANSSEN-CILAG	8751794
AbbVie	7910112
SANOFI AVENTIS FRANCE	7621548
GUERBET France	7587511
LABORATOIRE GLAXOSMITHKLINE	6975275
BRISTOL-MYERS SQUIBB	6340664
BOSTON SCIENTIFIC SAS	5880688
ETHICON	5618063
LES LABORATOIRES SERVIER	5518256
SORIN GROUP FRANCE	5453361

Table 3.2 – Entreprises dépensant le plus en cadeaux

Personnes	**Nombre d'entreprises touchées**
michallet mauricette	23
lambert marc	23
sibilia jean	22
milpied noel	21
vourzay catherine	21
berrebi alain	20
levy philippe	20
guerci bruno	19
richard philippe	19
durand philippe	19
krakowski ivan	19
compagnon florence	19
costa pierre	18
gayral monique	18
leclerc philippe	18

TABLE 3.3 – Personnes liées au plus grand nombre d'entreprises différentes

Personnes	**Montant (€)**
hanna khalil	309778
pinget michel	188229
cribier alain	144886
leclercq christophe	133764
laborde francois	132047
peyrin biroulet laurent	129310
marcellin patrick	124314
reynes jacques	124051
defaye pascal	123661
pierot laurent	120428
litzler pierre yves	119015
tchetche didier	112659
anselme frederic	111027
gras daniel	111007
flipo rene marc	108908

TABLE 3.4 – Personnes ayant reçu les plus gros volumes d'argent

Certaines irrégularités existent et seraient intéressantes à comprendre avec plus d'approfondissement. Ainsi on trouve une entreprise qui a donné à 1100 fois un cadeau de 144 € et parfois à une même personne via une succursale. Dans le même genre, une infirmière possède 294 liens avec une même entreprise pour un total de plus de 90000 €.

Entreprise	spécialité	montant
American Orthodontics	Orthopédie dentofaciale	42094
Zimmer Dental SAS	Chirurgie Orale	93970
SEPTODONT SAS	Médecine Bucco-Dentaire	308613
ROCHE SAS	Anatomie et cytologie pathologiques	286565
MSD France	Anesthesieréanimation	489946
bioMÃ©rieux	Biologie médicale	153502
MEDTRONIC France S.A.S	Cardiologie et maladies vasculaires	5106691
ETHICON	Chirurgie générale	461571
DEPUY France	Chirurgie maxillo-faciale	100995
GLOBAL D	Chirurgie maxillo-faciale et stomatologie	14229
DEPUY France	Chirurgie orthopédique et traumatologie	1307272
MEDTRONIC France S.A.S	Chirurgie infantile	53049
ALLERGAN FRANCE	Chirurgie plastique reconstructrice et esthétique	243234
SORIN GROUP FRANCE	Chirurgie thoracique et cardiovasculaire	1797789
JANSSEN-CILAG	Chirurgie urologique	747228
BARD FRANCE SAS	Chirurgie vasculaire	785254
ETHICON	Chirurgie viscérale et digestive	1648955
JANSSEN-CILAG	Dermatologie et vénéréologie	711887
SANOFI AVENTIS FRANCE	Endocrinologie et métabolisme	1258185
Genzyme	Génétique médicale	30869
NOVARTIS PHARMA SAS	Gériatrie	241254
MSD France	Gynécologie médicale	768905
FERRING SAS	Gynécologieobstétrique	516039
ROCHE SAS	Hématologie	933939
SANDOZ	Hématologie (option Maladie du sang)	49783
EUSA PHARMA SAS	Hématologie (option Onco-hématologie)	47829
GILEAD SCIENCES	Gastroentérologie et hépatologie	1862973
PFIZER SAS	Médecine du travail	40478
GILEAD SCIENCES	Qualifié en Médecine Générale	398134
MSD France	Médecine interne	566434
Genzyme	Médecine nucléaire	48415
ALLERGAN FRANCE	Médecine physique et réadaptation	361220
Fresenius Medical Care France S.A.S.	Néphrologie	909096
MEDTRONIC France S.A.S	Neurochirurgie	568522
BIOGEN FRANCE SAS	Neurologie	2298198
Fresenius Medical Care France S.A.S.	Neuropsychiatrie	21301
Zambon France	O.R.L et chirurgie cervico faciale	247396

FIGURE 3.3 – Entreprise la plus généreuse par spécialité - Première partie

LABORATOIRE GLAXOSMITHKLINE	Oncologie (option oncohématologie)	381552
ROCHE SAS	Oncologie option médicale	3039978
ELEKTA SAS	Oncologie option radiothérapie	66653
NOVARTIS PHARMA SAS	Ophtalmologie	2939480
Amplifon Groupe France	Oto-rhinolaryngologie	2062219
LABORATOIRE GLAXOSMITHKLINE	Pédiatrie	836638
VITALAIRE	Pneumologie	1505365
LES LABORATOIRES SERVIER	Psychiatrie	1407120
SHIRE France S.A.	Psychiatrie option enfant & adolescent	147065
GUERBET France	Radiodiagnostic	3432446
NOVARTIS PHARMA SAS	Radiothérapie	389253
ASTELLAS PHARMA	Réanimation médicale	81574
LES LABORATOIRES SERVIER	Recherche médicale	28981
PFIZER SAS	Rhumatologie	1224194
JANSSEN-CILAG	Santé publique et médecine sociale	41069
PIERRE FABRE MEDICAMENT	Stomatologie	19855
American Medical Systems France	Gynécoobstétrique et gynécologie médicale option 1	58449
ETHICON	Gynécoobstétrique et gynécologie médicale option 2	1247
THERABEL LUCIEN PHARMA	Spécialiste en Médecine Générale	69593
MSD France	Médecine Générale	1904380
GUERBET France	Radiodiagnostic et Radio- Thérapie	610353
MERZ PHARMA France	ORL et ophtalmologie	24662
MALLINCKRODT FRANCE SARL	Radiopharmacie	8617
ETHICON	Hygiène	34312
ASTELLAS PHARMA	Pharmacovigilance	51711
PIERRE FABRE MEDICAMENT	Hémovigilance	1956

FIGURE 3.4 – Entreprise la plus généreuse par spécialité - Seconde partie

Si certaines données plus précises permettent de mettre en évidence les liens entre certaines spécialités et des entreprises du secteur, comme le milieu dentaire, on ne peut pas pour autant en faire une généralisation. Ainsi même si dans les entreprises donnant le plus aux psychiatres on retrouve des fabricants d'antidépresseur, ce ne sont pas les plus gros budgets. En portant une plus grande attention à la nature des avantages, on se rend compte que ce lien apparaît en fait pour les spécialités nécessitant un matériel spécifique. L'avantage est alors un don de matériel.
Sur cette lignée, on peut distinguer deux profils d'avantages : les entreprises qui payent à un grand nombre de praticiens hébergement, transport et repas, et celles qui font don de matériel médical à un public plus restreint. Les premières semblent, au vue d'informations supplémentaires parfois renseignées, payer les frais pour des congrès ou autres événements liés au monde médical.

4 Conclusion

4.1 Perspectives futures

La loi Touraine passée le 26 janvier 2016 a amené un décret qui sera appliqué à partir du 1er Avril 2017. Ce décret vise à améliorer cette base en précisant le contenu des conventions signées entre corps médical et les industries. En précisant les rémunérations et avantages accordés dans le cadre de convention, la base de donnée permettra peut être de mettre plus facilement en évidence les influences des entreprises sur le milieu de la santé. En effet, les données que nous avons sont dédiées à des avantages normalement ponctuels plutôt qu'à une relation pérenne entre une entreprise et un praticien. Cependant, cela n'empêche pas certaines personnes de recevoir plusieurs dizaines de fois de la même entreprise.

Il y a donc possibilité pour un prochain projet d'IC05, de partir de notre travail et le lier avec un CSV des conventions commençant à partir d'avril.

4.2 Exploitation des données

A travers notre projet nous avons été confronté à une quantité importante de données. Visualiser toutes ces données n'était pas possible, c'est pourquoi nous les avons triées et nous n'avons choisi que de visualiser certaines d'entres elles. C'est la principale barrière que nous avons rencontré dans notre étude. Les outils tel que Neo4j nous ont permis de décortiquer petit à petit ces données mais malheureusement il nous est impossible de toutes les visualiser sur un même graphe.

Ce projet a été l'occasion pour nous de nous confronter au travail d'analyse de données et d'extraction de données pertinentes. Malheureusement à travers ce projet nous ne pouvons que formuler quelques hypothèses quant aux actions douteuses ou non de quelques entreprises ou professionnels de santé. Nous espérions par exemple trouver quelques noms de professionnels intervenant à la télévision ou auteurs de livres reconnus, mais ce ne sera pas le cas pour les avantages.