Projet Business Intelligence

Équipe Transparence Santé :

- BENALI Myriam
- BESSON Cécile
- CARDOSO MORENO Ineida
- NAAJI Dorian
- VIRARAGAVANE Smaïline

Introduction

Notre équipe de projet va travailler sur la base de données publique "Transparence santé".

La base de données publique Transparence - Santé rend accessible l'ensemble des informations déclarées par les entreprises sur les liens d'intérêts qu'elles entretiennent avec les acteurs du secteur de la santé. Pilotée par le ministère chargé de la santé, cette initiative de transparence vise à préserver la nécessaire relation de confiance entre les citoyens, les usagers et les multiples acteurs du système de santé.

Conformément à l'article L. 1453-1 du code de la santé publique, les entreprises produisant ou commercialisant des produits à finalité sanitaire ou cosmétique doivent rendre publics les avantages,les rémunérations accordés aux différents acteurs intervenant dans le champ de la santé, notamment aux professionnels de santé, ainsi que l'existence des conventions conclues avec ces acteurs.

Plus d'informations ici :

> Base de données transparence santé

> Téléchargement des données

Encodage erroné des données et minimisation

Encodage

Sur le site data.gouv, les données ont des problèmes d'encodages et les accents sont mal interprétés. L'ensemble des données devra être "nettoyé", via des routines basiques au sein de Talend afin de par exemple remplacer les caractères du genre :

> é = Ã©

> è = Ã¨

> à = Ã

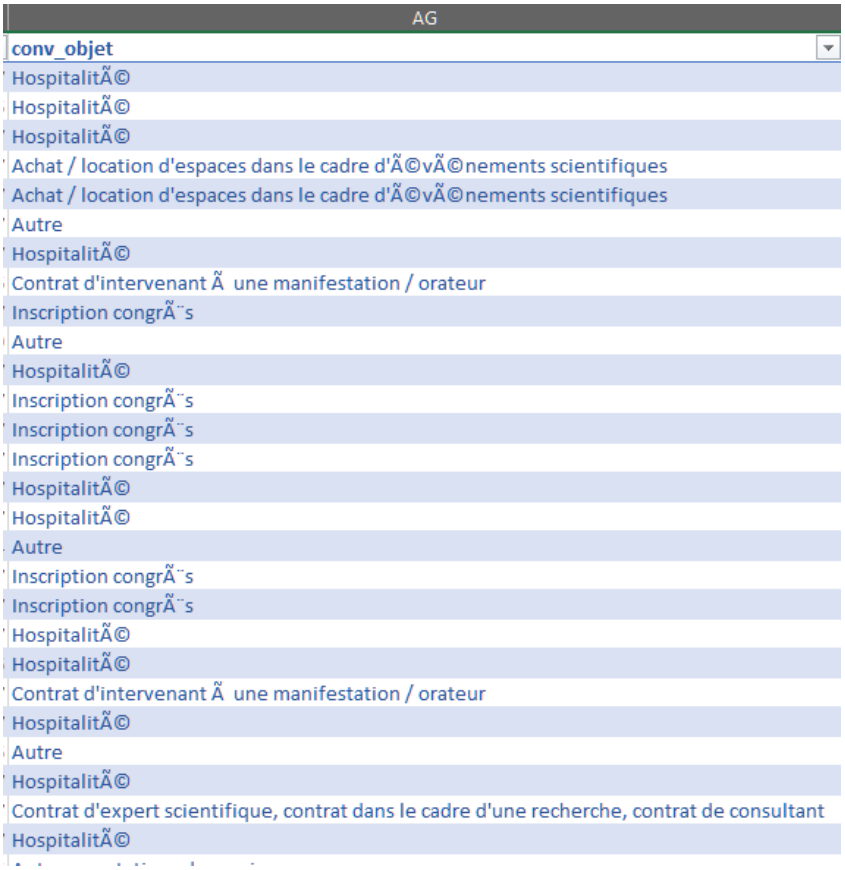
> Ã‰, Ã‰, Ãƒâ?° = É

> â? = '

> Å' = Œ

> Ã^, Ãƒ╦ = È

> etc. Trouver tous les caractères de ponctuation potentiellement mal encodés (majuscules comprises) et les remplacer par la suite (Talend).



AG
conv_objet
HospitalitÃ©
HospitalitÃ©
HospitalitÃ©
Achat / location d'espaces dans le cadre d'Ã©vÃ©nements scientifiques
Achat / location d'espaces dans le cadre d'Ã©vÃ©nements scientifiques
Autre
HospitalitÃ©
Contrat d'intervenant Ã une manifestation / orateur
Inscription congrÃ¨s
Autre
HospitalitÃ©
Inscription congrÃ¨s
Inscription congrÃ¨s
Inscription congrÃ¨s
HospitalitÃ©
HospitalitÃ©
Autre
Inscription congrÃ¨s
Inscription congrÃ¨s
HospitalitÃ©
HospitalitÃ©
Contrat d'intervenant Ã une manifestation / orateur
HospitalitÃ©
Autre
HospitalitÃ©
Contrat d'expert scientifique, contrat dans le cadre d'une recherche, contrat de consultant
HospitalitÃ©

Figure 1 : Mauvais encodage.

Voir : Tableau de correspondance de caractères "wrongly-encoded"

Minimisation

Le jeu de données fourni par data.gouv est très volumineux, d'autant plus une fois décompressé.

Volumétrie :

- declaration_avantage_2020_09_13_04_00 : `3.35 Go` (12 270 172 lignes)
- declaration_convention_2020_09_13_04_00 : `1.85 Go` (5 662 329 lignes)
- declaration_remuneration_2020_09_13_04_00 : `171 Mo` (614 365 lignes)
- entreprise_2020_09_13_04_00 : `398 Ko` (3316 lignes)

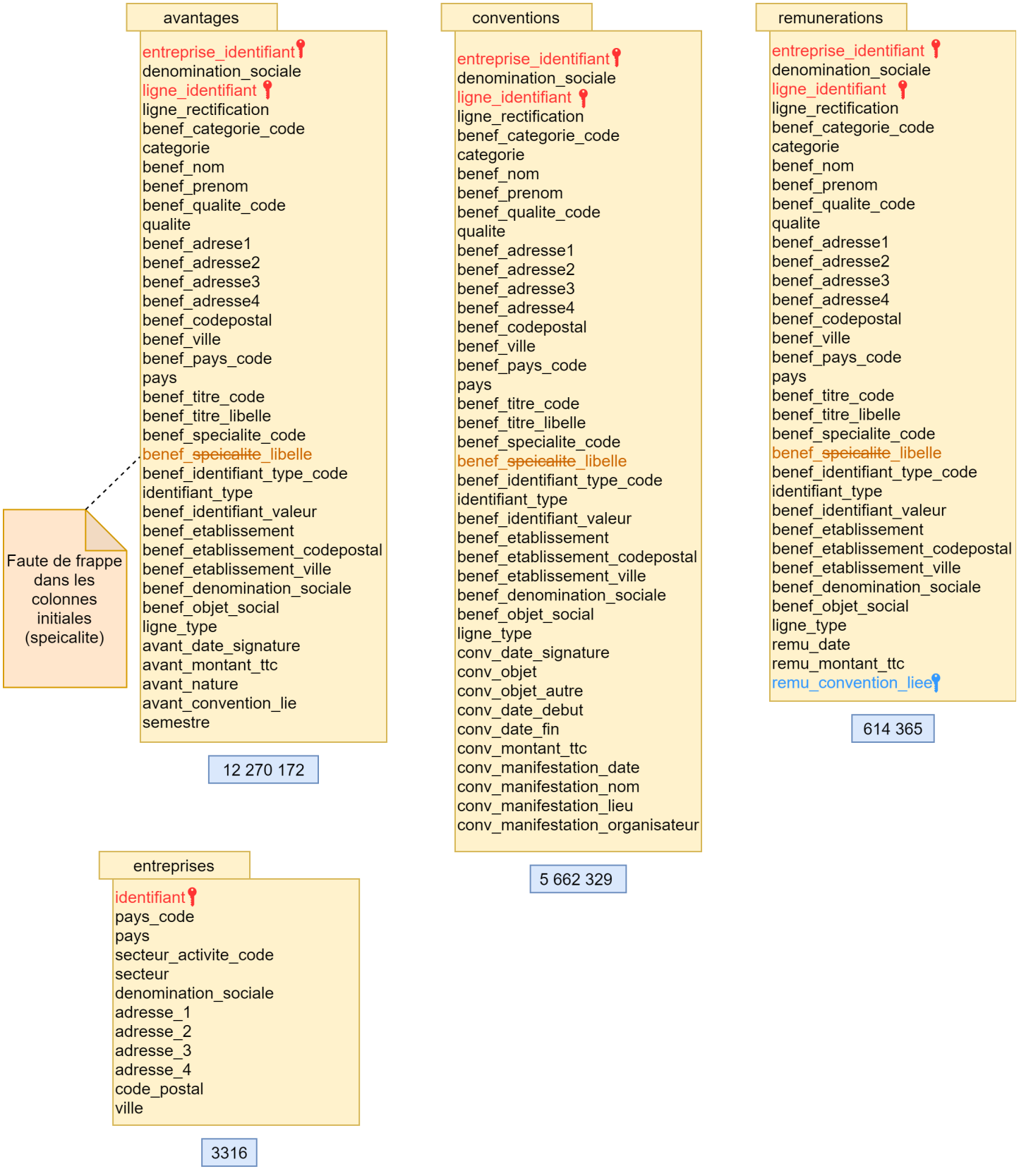


Figure 2 : Données initiales.

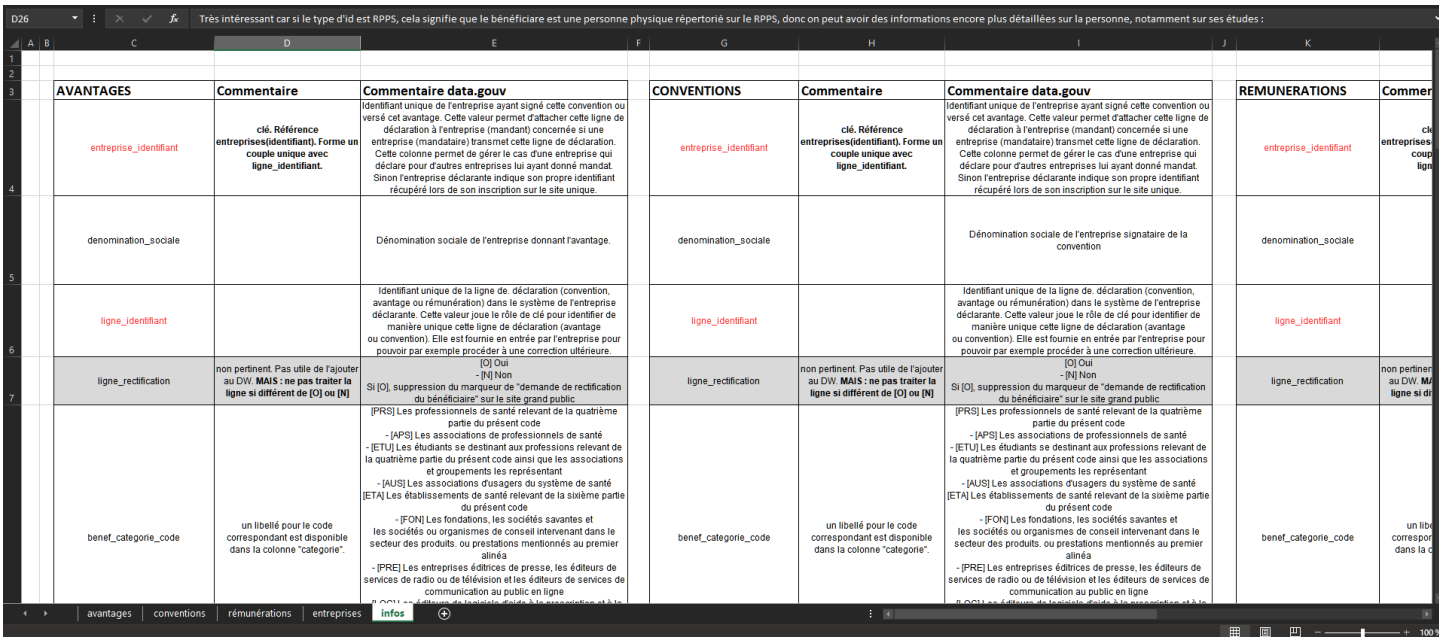
C'est pourquoi nous avons créé un jeu données minimisé, chacun gardant uniquement les "top 10 000 rows" de chaque fichier. Il permet d'ouvrir et de visualiser plus facilement les données. Ces données permettront également de réaliser des tests avant de passer à la totalité des données par exemple.

Ces données minimisées sont accessibles ici :

> Données minimisées (Accessible également sur le drive).

Un classeur Excel (format .xlsx) est également disponible ; il compile sous 4 feuilles l'ensemble des données minimisées, mises en forme de manière plus lisible, avec une page d'informations concernant les données.

> Classeur Excel (Accessible également sur le drive).



D26 — Très intéressant car si le type d'id est RPPS, cela signifie que le bénéficiaire est une personne physique répertorié sur le RPPS, donc on peut avoir des informations encore plus détaillées sur la personne, notamment sur ses études :

AVANTAGES	Commentaire	Commentaire data.gouv
entreprise_identifiant	**clé. Référence entreprises(identifiant). Forme un couple unique avec ligne_identifiant.**	Identifiant unique de l'entreprise ayant signé cette convention ou versé cet avantage. Cette valeur permet d'attacher cette ligne de déclaration à l'entreprise (mandant) concernée si une entreprise (mandataire) transmet cette ligne de déclaration. Cette colonne permet de gérer le cas d'une entreprise qui déclare pour d'autres entreprises lui ayant donné mandat. Sinon l'entreprise déclarante indique son propre identifiant récupéré lors de son inscription sur le site unique.
denomination_sociale		Dénomination sociale de l'entreprise donnant l'avantage.
ligne_identifiant		Identifiant unique de la ligne de. déclaration (convention, avantage ou rémunération) dans le système de l'entreprise déclarante. Cette valeur joue le rôle de clé pour identifier de manière unique cette ligne de déclaration (avantage ou convention). Elle est fournie en entrée par l'entreprise pour pouvoir par exemple procéder à une correction ultérieure.
ligne_rectification	non pertinent. Pas utile de l'ajouter au DW. **MAIS : ne pas traiter la ligne si différent de [O] ou [N]**	[O] Oui - [N] Non Si [O], suppression du marqueur de "demande de rectification du bénéficiaire" sur le site grand public
benef_categorie_code	un libellé pour le code correspondant est disponible dans la colonne "categorie".	[PRS] Les professionnels de santé relevant de la quatrième partie du présent code - [APS] Les associations de professionnels de santé - [ETU] Les étudiants se destinant aux professions relevant de la quatrième partie du présent code ainsi que les associations et groupements les représentant - [AUS] Les associations d'usagers du système de santé [ETA] Les établissements de santé relevant de la sixième partie du présent code - [FON] Les fondations, les sociétés savantes et les sociétés ou organismes de conseil intervenant dans le secteur des produits. ou prestations mentionnés au premier alinéa - [PRE] Les entreprises éditrices de presse, les éditeurs de services de radio ou de télévision et les éditeurs de services de communication au public en ligne

CONVENTIONS	Commentaire	Commentaire data.gouv
entreprise_identifiant	**clé. Référence entreprises(identifiant). Forme un couple unique avec ligne_identifiant.**	Identifiant unique de l'entreprise ayant signé cette convention ou versé cet avantage. Cette valeur permet d'attacher cette ligne de déclaration à l'entreprise (mandant) concernée si une entreprise (mandataire) transmet cette ligne de déclaration. Cette colonne permet de gérer le cas d'une entreprise qui déclare pour d'autres entreprises lui ayant donné mandat. Sinon l'entreprise déclarante indique son propre identifiant récupéré lors de son inscription sur le site unique.
denomination_sociale		Dénomination sociale de l'entreprise signataire de la convention
ligne_identifiant		Identifiant unique de la ligne de. déclaration (convention, avantage ou rémunération) dans le système de l'entreprise déclarante. Cette valeur joue le rôle de clé pour identifier de manière unique cette ligne de déclaration (avantage ou convention). Elle est fournie en entrée par l'entreprise pour pouvoir par exemple procéder à une correction ultérieure.
ligne_rectification	non pertinent. Pas utile de l'ajouter au DW. **MAIS : ne pas traiter la ligne si différent de [O] ou [N]**	[O] Oui - [N] Non Si [O], suppression du marqueur de "demande de rectification du bénéficiaire" sur le site grand public
benef_categorie_code	un libellé pour le code correspondant est disponible dans la colonne "categorie".	[PRS] Les professionnels de santé relevant de la quatrième partie du présent code - [APS] Les associations de professionnels de santé - [ETU] Les étudiants se destinant aux professions relevant de la quatrième partie du présent code ainsi que les associations et groupements les représentant - [AUS] Les associations d'usagers du système de santé [ETA] Les établissements de santé relevant de la sixième partie du présent code - [FON] Les fondations, les sociétés savantes et les sociétés ou organismes de conseil intervenant dans le secteur des produits. ou prestations mentionnés au premier alinéa - [PRE] Les entreprises éditrices de presse, les éditeurs de services de radio ou de télévision et les éditeurs de services de communication au public en ligne

REMUNERATIONS	Commer
entreprise_identifiant	cl… entreprises… coup… lign…
denomination_sociale	
ligne_identifiant	
ligne_rectification	non pertinen… au DW. MA… ligne si di…
benef_categorie_code	un libe… correspon… dans la c…

avantages | conventions | rémunérations | entreprises | infos

Figure 3 : Classeur excel.

Autres fournisseurs de données

D'autres sources de données peuvent nous être utile, notamment, nous avons trouvé plusieurs fichiers :

- dep_region.csv & communes-departement-region.csv : Issus de data.gouv ces fichiers peuvent nous aider à remplir la dimension des adresses.
- rpps_diplomes.csv : acessible sur annuaire.sante.fr, ce fichier nous permet, à partir d'un code unique attribué à un professionnel de santé (numéro RPPS) de récupérer des informations le concernant. Nous utiliserons un fichier permettant de connaître les études du professionnel de santé, nous permettant d'éventuelles analyses. Cela ne concerne que les bénéficiaires des avantages/conventions/rémunérations.

L'ensemble de ces .csv sont téléchargeables via ce lien gofile (lien disponible également sur le drive).

Entrepôt de données

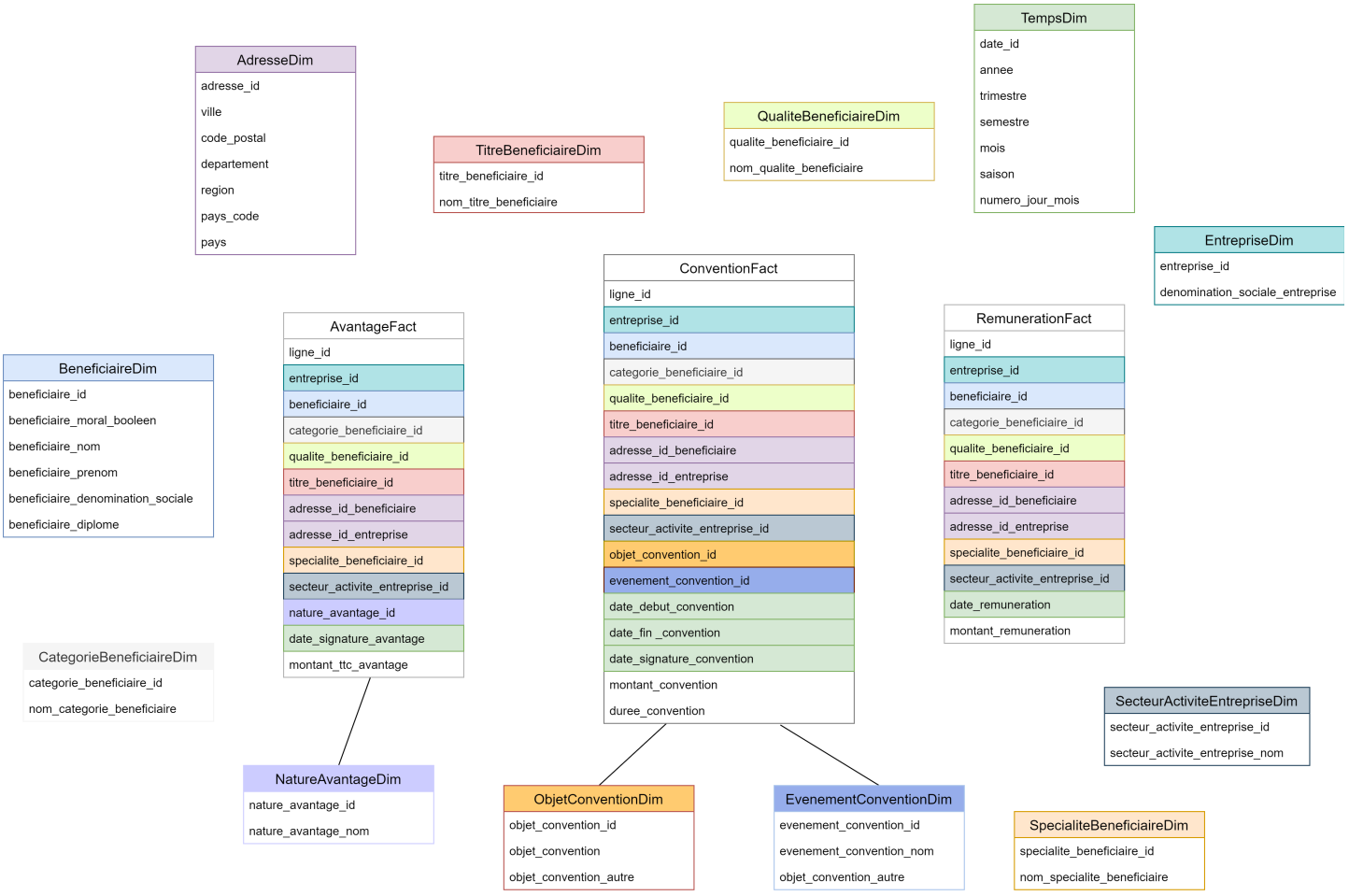


Figure 5 : Data Warehouse

La documentation complète et approfondie du Data Warehouse est accessible via ce lien.

Indicateurs clé de performance (KPIs)

Ces KPIs sont donnés à titre informatif et pourront évoluer au long du projet si de nouveaux KPIs pertinents sont trouvés.

Avantages

> Quelles sont les villes/pays où les entreprises font le plus appel à des acteurs du secteur de la santé en échange d'avantages ?

> Quelles sont les 5 premières professions de santé les plus demandées par les entreprises ?

> Quelles sont les 5 premières spécialités les plus recherchées par les entreprises ?

> Quel est en moyenne le montant des frais engendrés pour payer l'avantage au bénéficiaire ?

> Ce montant est-il plus ou moins élevé selon la ville, le poste, la spécialité du bénéficiaire ?

> Quels sont les principaux avantages accordés par les entreprises ? - Certaines entreprises font-elles plus appel à des prestataires que d'autre en échange d'avantages ?

> Quelle est le montant moyen des avantages en fonction des régions/départements/villes ?

Conventions

> Quelle est la durée moyenne d'une convention ?

> Certaines entreprises établissent-elle significativement plus de conventions que d'autres ?

> Y-t-il un lien entre le nombre de convention établies entre une entreprise et une prestation selon la ville ou la spécialité ?

> Quel est le montant moyen d'une convention ? En fonction des villes, régions, départements ?

> Le moment de l'année influence-t-il la signature des conventions ?

> De quel secteur sont issues les entreprises signant le plus de conventions ? Nombre de conventions signées en fonction du secteur d'entreprise.

> Quels sont les objets de convention les plus récurrents ? Les plus chers ?

> Zone géographique des bénéficiaires/entreprises en fonction du montant moyen de la convention

Rémunérations

> Quel est le montant moyen d'une rémunération ?

> Certains professionnels sont-ils mieux rémunérés que d'autres selon la ville, leur poste, leur spécialité ?

> Certaines entreprises font-elles plus appel à des prestataires que d'autres en échange d'une rémunération ?

> Quelle est la rémunération moyenne en fonction des régions/départements/villes ?

> Quel secteur d'entreprise fournit les plus grosses rémunérations ? les plus faibles ? montant moyen des rémunérations en fonction du secteur ? en fonction de la qualité du bénéficiaire ?

> La date / moment de l'année / saison influence-t-elle le montant des rémunérations ?

Gestion de projet :

Un drive dédié au partage des différents documents et données liés au projet est disponible à l'adresse suivante

Nous gérons aussi nos tâches via un formalisme simple via Trello, un outil en ligne de collaboration.

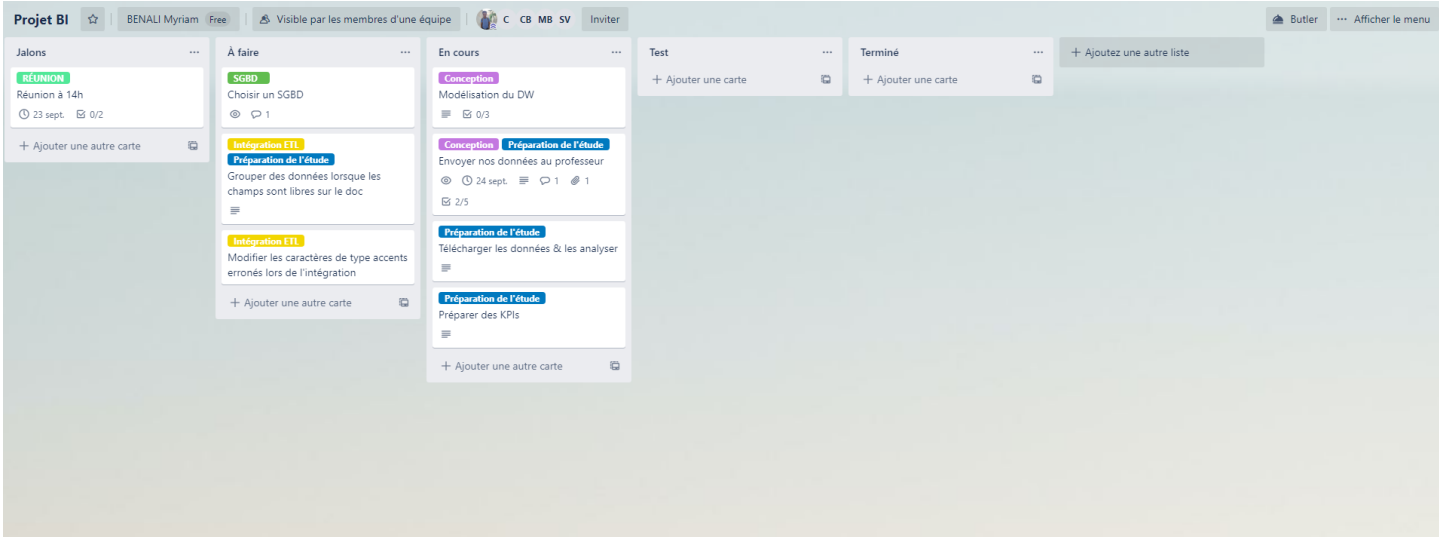


Figure 5 : Organisation des tâches.

Il est accessible ici (click me).