Datasets de données RSE sociétaux des entreprises

13/12/2021

—

Totally Sprint

Benmesmoudi Nadjelaa

Pamidjo Jocelyne

Nguy Sioutyne

Diallo Coumba

Chapitre 8: Présenter les données (data paper)

Présentation du Sprint

- objectifs

Réussir à avoir la base de données finale qui regroupe nos précédents jeux de données avec une architecture bien définie.

- acteurs

Les acteurs de ce sprint sont Sioutyne NGUY comme scrum master, Nadjelaa BENMESMOUDI comme project owner. Et tous les membres du groupe, Coumba DIALLO, Jocelyne TRAORE, Nadjelaa BENMESMOUDI et Sioutyne NGUY en qualité de développeurs.

- évènements

Finaliser la base de donnée finale et pouvoir la stocker sur un espace de stockage

- support et outils utilisés

Python, excel et trello

Tâches réalisées

- Résumé

La responsabilité sociétale et environnementale des entreprises comporte de nombreux enjeux pour l'entreprise qui souhaite intégrer sa stratégie globale, liés aux objectifs de développement durable.

La RSE comporte un volet social important qui se traduit par des objectifs en matière de conditions de travail, de bien-être et de motivations des collaborateurs. Les actions mises en place visent à améliorer l'environnement de travail et augmenter la satisfaction des collaborateurs. Dans notre étude, nous nous sommes intéressés plus précisément aux inégalités hommes-femmes dans le milieu professionnel en termes d'effectif, rémunération et plein d'autres enjeux.

- Contexte et objectifs

Durant ce cours, nous avons dû créer des équipes et définir le rôle de chacun afin de gérer un projet avec la méthode agile. Ainsi, nous voilà une équipe de 5 personnes entièrement féminine et dont le nom est Totally Sprint.

Ce nom fait référence aux Totally Spies, un dessin animé de notre enfance ou des filles sauvent le monde grâce à des missions ! Fantastique n'est-ce pas ?

A partir de métadonnées fournies par notre professeur, données portant sur la RSE des entreprises. Nous avions comme objectif de chercher et sélectionner des indicateurs pertinents afin de mettre en avant la responsabilité sociétale des entreprises. De plus, grâce à notre étude, nous avons la possibilité de mettre en avant les bons comme les mauvais comportements des entreprises. Notre étude pourra aider les entreprises à s'améliorer et à évoluer pour le bien de la société.

Etant une équipe de filles, le sujet de l'égalité salariale hommes femmes au travail nous paraissait le plus important comme nous y sommes actuellement et dans le futur impactés. De plus, nous sommes tous concernés par l'environnement et par conséquent l'empreinte carbone des entreprises.

- démarche et organisation de la base de données

  La récolte de données RSE sociétaux est souvent liée à la recherche d'indicateurs pariétaux ou encore à ceux affiliés à des indicateurs comme l'égalité, à une géolocalisation des structures ou encore à des données liées à des critères beaucoup plus spécifiques.Ainsi dans le cadre de la mise en place de nos jeux de données, certains jeux de données ont été jugés pertinents pour nos recherches pour aboutir à l'établissement d'un jeu de données final.Ainsi plusieurs étapes successives sont notées dans ce processus:

  - Collecte des jeux de données sur différentes sources notamment Datagouv et INSEE
  - L'identification des tables et le choix des différents attributs des tables
  - Le nettoyage des datasets
  - La mise en place d'un schéma relationnel pour avoir une structuration du dataset final
  - Procéder à une jointure de toutes les différentes tables

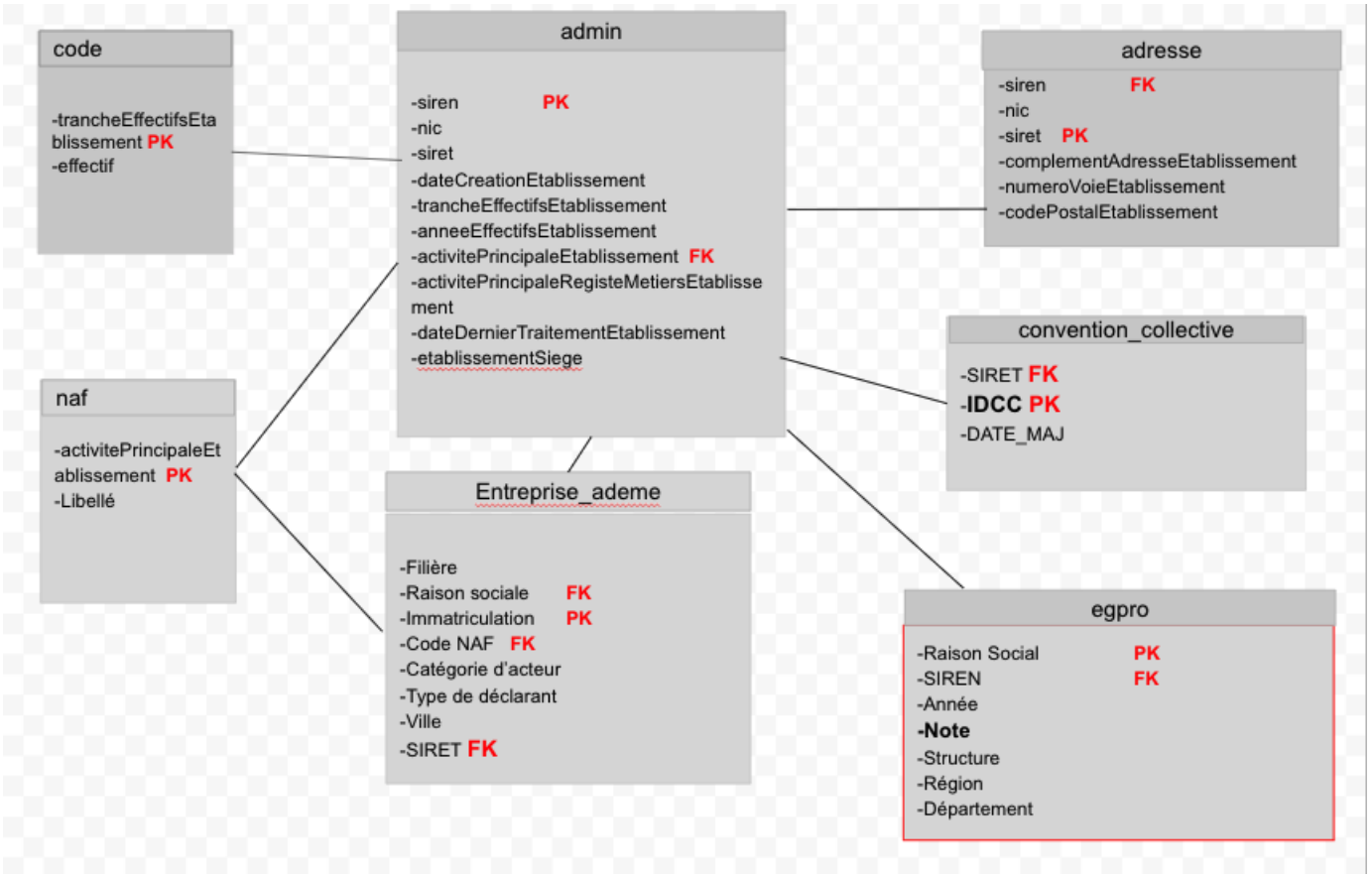
- Obtention de la table finale
- Mise à disposition des annexes sur Gitea

Organisation du dataset

Le dataset final est une jointure des différents datasets que nous avons eu à collecter:

Adresse

- Description des variables



Variable	Type	Source/Ingénierie
Siren	numeric	admin
nic	numeric	admin
siret	numeric	admin
dateCreationEtablissement	date	admin
trancheEffectifsEtablissement	numeric	admin
anneeEffectifsEtablissement	integer	admin
activitePrincipaleEtablissement	character	naf
Libellé	character	admin
activitePrincipaleREgistreMetiersEtablissement	character	admin
dateDernierTraitementEtablissement	date	admin
etablissementSiege	character	admin
complementAdresseEtablissement	character	adresse
numeroVoieEtablissement	numeric	adresse
CodePostaleEtablissement	numeric	adresse
trancheEffectifsEtablissement	numeric	code
effectif	numeric	code
Raison Social	character	egpro
Année	numeric	egpro

Note	numeric	egpro
Structure	character	egpro
Région	character	egpro
Département	character	egpro
IDCC	numeric	convention_collective
DATE_MAJ	date	convention_collective
Filière	character	Entreprise_ademe
Immatriculation	numeric	Entreprise_ademe
Catégorie d'acteur	character	Entreprise_ademe
Type de déclarant	character	Entreprise_ademe
Ville	character	Entreprise_ademe

- Description de chaque table source:

1.egpro :

Description

L'Index de l'égalité professionnelle a été conçue pour faire progresser au sein des entreprises l'égalité salariale entre les femmes et les hommes.

Il permet aux entreprises de mesurer, en toute transparence, les écarts de rémunération entre les sexes et de mettre en évidence leurs points de

progression. Lorsque des disparités salariales sont constatées, des mesures de correction doivent être prises.

L'index est calculé à partir de 5 indicateurs portant sur l'écart des rémunérations, promotions et augmentations, mais également sur le congé maternité et les plus hauts salaires de l'entreprise.

2.convention_collective :

Description

Ce jeu de données présente les conventions collectives déclarées par entreprise (SIRET)

Ce jeu de données est exploité dans le projet Code du travail numérique

Il est issu de la DSN et mis à jour ponctuellement avec plusieurs mois de retard.

Le colonne IDCC représente le numéro de convention collective, laquelle peut être consultée via https://beta.legifrance.gouv.fr/recherche

3.Entreprise_ademe (REP-EMB):

Description

Le jeu de données

Le jeu de données présenté ici comprend la liste des entreprises de la filière "Emballages ménagers" (EMB) inscrites dans SYDEREP.

Date des données : juin 2020 (en cours d'actualisation)

Règles d'actualisation des données

Les données concernent la liste des producteurs enregistrés à un instant T sur SYDEREP, soit par l'intermédiaire d'un éco-organisme, soit en direct dans le cas d'un système individuel. Ces listes ne reflètent pas les producteurs qui ont effectué une déclaration sur SYDEREP.

L'actualisation de cette liste est mise à disposition une fois par an sur https://data.ademe.fr.

Généralités sur les filières REP

L'État a confié à l'ADEME la mise en place et la gestion des registres de déclarations annuelles obligatoires des metteurs sur le marché relevant des REP suivantes :

Déchets des équipements électriques et électroniques (DEEE) ;

Piles et accumulateurs (PA) ;

Pneumatiques usagés (PU) ;

Véhicules automobiles hors d'usage (VHU) ;

Déchets d'éléments d'ameublement (DEA) ;

Papiers graphiques (PAP) ;

Emballages ménagers (EMB) ;

Déchets issus de bateaux de plaisance ou de sport (DBPS).

Cette liste va s'enrichir au fur et à mesure de l'entrée en vigueur de la loi AGEC avec :

Au 01/01/2022, les filières : Produits du tabac, Lubrifiants (HU), Articles de bricolage et de jardin (ABJ), Articles de sport et loisirs (ASL), Jouets, Produits et matériaux du secteur de la construction du bâtiment (PMCB),

Au 01/01/2023, la filière Emballages issus de la restauration,

Au 01/01/2024, les filières Gommes à mâcher et Textiles sanitaires à usage unique,

Au 01/01/2025, la filière Engins de pêche

Ces registres sont gérés dans l'application SYDEREP, permettant aux éco-organismes ou aux producteurs en système individuel, le renseignement direct de leurs données. Ces registres ayant pour objet la mesure de la performance globale au niveau national de chacune des filières REP, l'ADEME établit et publie annuellement, via sa Librairie un état des lieux des filières avec un certain nombre d'indicateurs agglomérés, complétés d'éléments d'analyses qualitatives provenant des acteurs de chaque filière.

- Exploitation et usages

Nous disposions dès le départ d'informations sur les entreprises à partir d'une base de données comportant le numéro SIREN / SIRET, le code NAF , les tranches d'effectif et plusieurs éléments liées à l'identification des entreprises et des établissements.

L'enjeu pour nous était donc de pouvoir exploiter ses informations en trouvant d'autres données qu'on pourrait croiser avec celles que nous avons déjà afin de déduire des informations pertinentes et utiles . La mise en commun des données est

essentielle pour permettre de développer une stratégie data à haute valeur ajoutée pour l'entreprise.

En croisant ces différentes données, il est possible d’accéder à des informations plus pointues et surtout plus riches sur leur marché, leurs métiers et surtout leur situation face aux problématiques actuelles.

Dans notre cas nous nous sommes beaucoup intéressés aux sujets liées à l'égalité professionnelle , l'écart des rémunérations, promotions et augmentations entre homme et femme. Et surtout la gestion des déchets générés par les produits fabriqués ou mis sur le marché (responsabilité élargie du producteur REP).

Notre base de données permettra donc à connaître pour chaque entreprise sa situation face au traitement entre les femmes et les hommes en termes d'accès à l'emploi, à la formation, à la mobilité et à la promotion ou en termes d'égalité salariale. Si elle arrive à gérer correctement ses résidus de fabrication de produits, si elle participe à réduire les déchets qui polluent la planète et si elle est exempte de toute réprimande.

Le plus important est en fait de mettre en évidence les points de progression sur lesquels agir. La finalité est qu'à partir de nos extractions qu’on puisse avertir une entreprise sur sa façon de fonctionner et qu'elle puisse prendre des actions afin d'éviter de mauvaises notations

Difficultés rencontrées

- ☐ Volume de données
- ☐ Trouver des jeux de données facile à intégrer
- ☐ Temps limité
- ☐ Importer les données sur système de gestion de BDD (mysql, postgres)

Ce qu'il faut retenir

- Qu'est-ce qu'un data paper ?

Le data paper est une publication qui décrit un jeu de données scientifiques brutes (data, dataset), notamment à l’aide d’informations précises, appelées métadonnées (metadata).

Il est publié sous forme d'articles examinés par les pairs dans une revue publiant généralement différentes formes d'articles dont des data papers ou dans un data journal, c'est-à-dire une revue contenant exclusivement des data papers.

Le data paper comprend deux parties :

- l'ensemble des fichiers des données (data files) accessibles directement ou via un entrepôt de données. Ces données sont en libre accès ou en restriction d'accès temporaire ;

- La partie descriptive, c'est le data paper proprement dit. Cette partie explique le contexte d'obtention des données, les présente et en démontre la fiabilité.

Le data paper a pour objectif d'informer la communauté scientifique de l'existence et de la disponibilité d'un jeu de données qui est déposé dans un entrepôt de données. Dans notre cas nous utiliserons comme canal Gitea :

Il valorise les données en exposant leur potentiel pour des utilisations et projets futurs. Il facilite leur réutilisation en mettant en évidence la qualité des données et des procédures, ainsi que la rigueur scientifique de l'étude. Il apporte de la visibilité aux données, les rend plus facilement repérables et citables par d'autres études. Ainsi notre travail pourra être utile à d'autres personnes en recherche d'information tel que l'index d'égalité professionnelle par exemple sur une entreprise ou un établissement français à partir du numéro SIREN. Le data paper met aussi en valeur ses auteurs en tant que créateurs de données.

Notre dépôt complet sur Gitea ICI