RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

Dares
Déchiffrer le monde du travail
pour **éclairer** le débat public

Base Jocas

DESCRIPTION DES VARIABLES DE LA BASE JOCAS

Table des matières

Identifiants de l'offre

url

url de l'offre scrapée. L'url et la date de scraping sont utilisées comme identifiant des pages html scrapées et stockées par la Dares. Certains sites sont scrapés en faisant directement appel à une API (Apec, Api de Pôle emploi, Météojob). Pour ces sites l'url est renseignée, mais il n'y a pas de page html téléchargées : nous enregistrons le résultat des requêtes.

ID_JOCAS

Identifiant que nous attribuons à une offre. Cet identifiant est supposé unique (*en pratique il y a un overlap des ID pour Pôle emploi début 2019 pour une raison que j'ignore*). Les identifiants sont formés de la manière suivante :

Nom du site + '_ ' + date_firstSeenDay + '_' + Numéro de l'offre pour le jour de scraping

Metadata du scraping

date_scraping

Date et heure à laquelle l'offre a été scrapée (où à laquelle on a tenté de la scraper si échec du scraping). Ici, scrapée signifie téléchargée sur le site d'emploi.

Format : %a %b %d %H:%M:%S %Y (Exemple : Fri Jun 4 22:20:56 2021)

scrapingFailure_status

Booléen qui indique si l'offre a été correctement téléchargée ou non (True : le scraping a échoué ; False : le sraping a réussi). Les offres peuvent ne pas être correctement scrapées en cas de problème de connexion (côté client comme côté serveur), d'interdiction d'accès à l'offre, de retrait de l'offre du site entre le moment où on la détecte et celui où on la télécharge…

Sites et partenaires

site_name

Nom du site. Selon les périodes, les sites scrapés peuvent varier (car nous en ajoutons et que certains sites disparaissent).

Site_name	Début de la collecte	Fin de la collecte
api_poleemploi	23/01/2019	
regionsjob	01/01/2019	
jobintree	01/01/2019	
cadremploi	15/07/2020	
bdm	22/10/2019	15/12/2020
apec	01/03/2019	
cadreo	22/10/2019	
viadeo	01/01/2019	29/06/2019
leboncoin	01/01/2019	
keljob	15/07/2020	
meteojob	14/12/2020	

site_child

Nom du sous-site (si applicable). Le site Régionsjob se divisent en plusieurs sous-sites régionaux : Centre, Est, Nord, Ouest, Paca, Paris, RhoneAlpes, SudOuest.

partner_status

Booléen indiquant si l'offre provient d'un site partenaire ou non. Les sites se redirigent les uns les autres. Nous essayons d'identifier les redirections entre sites, quand elles sont explicitement indiquées.

partner_name

Nom du site partenaire (si applicable). Les sites partenaires sont parfois (mais pas toujours) des sites que nous scrapons par ailleurs.

Dates

Pour cette partie, le « jour » correspond au jour de lancement du scraper, cf. date_firstSeenDay

date_firstSeenDay

Jour où l'offre a été vue sur le site pour la première fois. Ce jour ne correspond pas forcément au jour de scraping de l'offre, c'est le jour du lancement du scraper. Par exemple, si le scraper est lancé le lundi soir mais que l'offre est scrapée plus tard lors de la session de scraping, par exemple le mardi à 00h30, la valeur de « date_firstSeenDay » sera le lundi et non le mardi. Attention, cela n'est pas forcément vrai pour les offres scrapées en 2019 et début 2020 (ie avant la mise à jour des scrapers).

Format : %Y-%m-%d

date_firstDisappearedDay

Premier jour où l'offre n'a plus été vue sur le site.

Format : %Y-%m-%d

date_lastSeenDay

Jour où l'offre a été vue sur le site pour la dernière fois. Cela ne correspond pas toujours à la veille de « date_firstDisappearedDay » notamment en cas de bug des scrapers. S'il n'y a pas eu de discontinuité dans le scraping, c'est bien la veille du premier jour où l'offre n'est plus vue.

Format : %Y-%m-%d

date_sitePublicationDay

Date de publication de l'offre telle que renseignée sur le site. C'est une information fournie par le site contrairement aux variables précédentes. Ce champ peut être vide.
Format : %Y-%m-%d

Métier et qualification

job_title

Intitulé (ou titre) de l'offre d'emploi, tel que renseigné sur le site.

job_ROME_code

Code métier selon la nomenclature ROME de Pôle emploi. Le code ROME est codé automatiquement à partir de l'intitulé de l'offre (variable « job_title »), grâce à un algorithme décrit dans ce document. Pour l'API Pôle emploi, les offres contiennent déjà le code ROME, codé par Pôle emploi.

job_qualification

Qualification du métier, selon les niveaux définis par la Dares :

- 0 Indéterminé ou non renseigné
- 2 Manoeuvre et ouvrier non qualifié
- 4 Ouvrier qualifié et ouvrier hautement qualifié
- 6 Employé non qualifié et employé qualifié
- 7 ou 8 Technicien, agent de maîtrise et assimilé
- 9 Ingénieur et cadre

Seule les offres de l'API Pôle emploi et de l'Apec contiennent le niveau de qualification. Pour l'instant cette variable n'est pas recodée sur les autres sites.

Contrat

contractType

Type de contrat proposé, tel que renseigné dans les champs structurés des offres. Les catégories ne sont pas homogènes selon les sites. Dans certains cas nous pouvons pas distinguer entre plusieurs contrats. Les differentes valeurs possibles sont : CDI ; CDD ; MIS [Mission d'Intérim] ; Apprentissage ; Alternance ; Independant ; Independant/Franchise ; Stage/Alternance ; Stage ; Apprentissage/Alternance ; CDD/MIS ; CDI/MIS ; Franchise ; Titulaire de la fonction publique ; Contrat étranger ; Contrat de professionnalisation ; VIE ; VIA ; Bénévolat ; CDS [CDD Sénior] ; SAI [Saisonnier] ; CCE [Profession commerciale]; REP [Reprise d'entreprise] ; TTI [Contrat travail temporaire insertion] ; DDI [Contrat durée déterminée insertion] ; DIN [CDI Intérimaire] ; Intermittent ; Contrat d'usage ;Contractuel de la fonction publique ; Avis de concours

contractDuration_min

Durée minimale du contrat. Cette variable n'est pas disponible dans un champ structuré pour Keljob et Viadeo.

contractDuration_max

Durée maximale du contrat. Si la durée du contrat est fixe, cette variables est vide.Cette variable n'est pas disponible dans un champ structuré pour Keljob et Viadeo.

contractDuration_period

Période dans laquelle la durée du contrat est exprimée. Les catégories possibles sont : DAY ; WEEK ; MONTH ; YEAR

contractDuration_value

Durée moyenne du contrat, exprimée en jours.

Temps de travail

workTime_hours

Temps de travail exprimée en heures. Cette variable est seulement disponible pour l'API Pole Emploi.

workTime_category

Temps de travail, selon si c'est un temps plein ou un temps partiel. Cette variable est disponible comme un champ structuré pour l'Apec, BDM, Cadreo, Leboncoin, et Regionsjob. Pour l'API Pôle Emploi, cette variables est inférée selon la variable *workTime_hours*, et le suffixe _INFER est ajouté pour les identifiés.

Les valeurs possibles sont : FULL_TIME ; PART_TIME ; FULL_TIME_INFER ; PART_TIME_INFER

workTime_value

Temps de travail exprimé en heures. Pour l'API Pole Emploi, elle reprend le nombre d'heures exprimé dans workTime_hours. Pour les autres sites, elle reprend l'équivalent horaire du temps partiel ou temps complet, 35 ou 24 respectivement.

Description

Attention, les variables description_job, description_profil et description_entreprise sont expérimentales et peuvent être de mauvaise qualité (mauvais parsing de la description).

description_full

Champ libre non structuré, qui correspond au corps du texte de l'offre d'emploi. Ce champ peut contenir des caractères spéciaux, des balises html, des smileys… Il peut y avoir des problèmes d'encoding. Sur certains sites, la description est divisée en plusieurs partie : la description du poste, le profil attendu du candidat et une description de l'entreprise. Lorsque c'est possible nous identifions ces parties (notamment grâce aux balises html).

description_job

Description du poste.

description_profil

Profil du/de la candidat.e idéal.e.

description_entreprise

Description de l'entreprise.

Lieu de travail

location_label

Libellé le plus récis du lieu de travail, tel que renseigné par le site. Cela correspond généralement à la commune, mais cela peut être une région ou un pays.

location_zipcode

Code postal du lieu de travail. C'est le code postal indiqué par le site, si cette donnée est disponible. Sinon la Dares effectue un matching entre « location_label » et les libellés de communes françaises pour retrouver le code postal. Parfois le matching est ambigüe (plusieurs codes postaux pour un même libellé), dans ce cas le code postal reste vide.

location_departement

Département du lieu de travail (départements français ou DOM-TOM). Le département « 99 » correspond à l'étranger.

location_country

Si l'offre est à l'étranger et que le pays étranger est connu, nom du pays étranger. Cette variable est généralement vide car la majorité des offres sont en France.

Salaire

salary_min
Salaire minimum proposé dans l'offre.

salary_max
Salaire maximum proposé dans l'offre. Cette variable est vide si le salaire est fixe.

salary_period
Plage temporelle dans laquelle le salaire est exprimé. Si la plage temporelle n'est pas disponible, celle-ci est inferée selon la méthodologie décrite dans le document méthodologique, et le suffixe _INFER est ajouté pour les indentifier.

Les valeurs possibles sont : DAY, DAY_INFER, HOUR, HOUR_INFER, MONTH, MONTH_INFER, WEEK, YEAR, YEAR_INFER.

salary_hourly_mean
Salaire horaire moyen. Calcul selon la métoholodie dans le document methodologique.

salary_hourly_min
Salaire minimum proposé converti en salairehoraire.

salary_max
Salaire maximum proposé dans l'offre converti en salaire horaire. Cette variable est vide si le salaire est fixe.

Entreprise

entreprise_nom

Nom de l'entreprise (renseigné par le site, brut sans nettoyage).

entreprise_siren

Numéro Siren de l'entreprise. Pour l'instant, seul un site (Leboncoin) contient le numéro Siren. A terme ce champ sera codé par la Dares.

Secteur d’activité

Le secteur d’activité est renseigné en NAF pour l’API Pôle emploi et l’Apec et sera codé par la Dares pour les autres sites. Certains sites précisent un « secteur d’activité » selon leur nomenclature propre : il ne s’agit pas de cela.

entrepriseSecteur_NAF21

Secteur d’activité de l’entreprise selon la nomenclature NAF de l’Insee, au niveau A21.

entrepriseSecteur_NAF88

Secteur d’activité de l’entreprise selon la nomenclature NAF de l’Insee, au niveau A88.

Télétravail

teleworking_accepted

Booléen permetant d'identifier si le télétravail est accepté ou non. Cette variable est codée pour Apec, Méteojob et RegionsJob qui ont des champs specifiques sur le télétravail, et si dans la description est presente la formulation « télétravail : non ».

teleworking_type

Type de télétravail accepté. Cette variables est disponible pour l'Apec et Météojob.

Les valeurs possibles sont : Partial, Total.

teleworking_mentioned

Booléen permetant d'identifier si dans la variable description_full sont mentionés les mots « télétravail », « travail à distance » ou « travail à la maison ».

Expérience

experience_min

Variable numérique correspondant au nombre minimum d'années d'expérience nécessaire pour le poste (0 s'il est précisé que les personnes sans expérience sont acceptées).

experience_max

Variable numérique correspondant au nombre maximum d'années d'expérience nécessaire pour le poste. Vide si seule l'expérience minimale est spécifiée dans l'offre.

Education

education_level

Variable sous forme de liste contenant les différents niveaux d'éducation / diplômes compatibles avec le poste, parmi :

- **CAP, BEP et équivalents**: Include CAP, BEP and GCSE levels.
- **Bac ou equivalent**: A levels or high school diploma. includes technical and professional baccalaureate
- **Bac+2 ou équivalents**: Diploma of Higher Education or Associate's Degree
- **Bac+3, Bac+4 ou équivalents**: Bachelor and first year of master
- **Bac+5 et plus ou équivalents**: Master's degree and Phd.

education_field

Variable textuelle correspondant au domaine d'études dans lequel le candidat doit posséder un diplôme. L'information n'est disponible directement que pour un site. Pour les autres sites, si un diplôme spécifique à un domaine est précisé (ex : BAFA), il est renseigné ici.