RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

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

BASE JOCAS

Document méthodologique résumé

Depuis fin 2018, la Dares collecte quotidiennement des offres d'emploi publiées en ligne sur un échantillon de sites de recrutement. Cette collecte massive a permis la création d'une nouvelle base de données composée du contenu détaillé de plusieurs millions d'offres d'emploi chaque année : Jocas (*Job offers collection and analysis system*).

Contexte

Ces dernières années, la part des annonces d'offre d'emploi ayant fait l'objet d'une publication en ligne a fortement augmenté : d'après l'enquête sur les offres d'emploi et les recrutements (Ofer) menée par la Dares, cette part est passée en France de 53 % en 2005 (BESSY et MARCHAL, 2006) à 95 % en 2016 (BERGEAT et REMY, 2017). La publication en ligne des offres d'emploi permet notamment d'accéder facilement à leur contenu détaillé (titre de l'offre, description, salaire, qualification…) et ouvre ainsi un nouveau champ d'exploitation de ce type de données, qui doit permettre d'améliorer la compréhension du marché du travail. Les offres d'emploi en ligne ont aussi l'avantage de refléter « en temps réel » la situation sur le marché du travail, ce qui peut permet d'anticiper certaines tendances avant la publication de chiffres officiels issus des sources administratives ou de résultats d'enquêtes. La Dares a ainsi décidé de collecter quotidiennement les offres d'emploi en ligne publiées sur une quinzaine de sites afin de stocker ces informations dans une base de données. La base Jocas vient ainsi s'ajouter aux sources de données habituellement utilisées sur les offres d'emploi et les emplois vacants, qu'elles soient administratives (STMT, DPAE, MMO) ou issues d'enquêtes statistiques (Acemo, BMO, EEC).

Aujourd'hui, les offres d'emploi sur Internet sont ainsi principalement accessibles sur :
- Les sites d'emploi « propriétaires », comme les sites « carrière » des entreprises et les sites d'agences de recrutement ou d'intérim, qui diffusent leurs propres offres.
- Les *jobboards*, qui sont des intermédiaires directs entre candidats et recruteurs. Ils diffusent auprès des candidats les offres d'emploi publiées par les recruteurs sur le site.
- Les agrégateurs, qui sont des moteurs de recherche d'offres d'emploi. Ils indexent des *jobboards* ou des sites « propriétaires » afin de proposer au candidat un large panel d'offres d'emploi. Si un candidat veut postuler, il est redirigé vers le site d'origine de l'offre.
- Les réseaux sociaux exclusivement dédiés à l'emploi comme LinkedIn et certains réseaux sociaux généralistes.
- Les médias en ligne spécialisés proposant une rubrique pour le partage d'annonces (le site de « L'étudiant » dispose par exemple d'une page dédiée aux offres d'emploi).

Les frontières entre les différents types de sites d'emploi sont toutefois assez poreuses.
D'une part, il semble qu'aujourd'hui la plupart des agrégateurs ne se contentent plus d'indexer des contenus existants mais permettent aussi aux recruteurs de déposer des offres sur leur site et aux candidats d'y postuler.

D'autre part, certains *jobboards* traditionnels diffusent désormais des offres provenant de sites tiers. C'est notamment le cas de Pôle emploi, qui, en plus de diffuser l'ensemble des offres déposées à Pôle emploi sur le site pole-emploi.fr, publie depuis 2013 sur son site des offres provenant de sites d'emploi partenaires.

Bien que le paysage des sites d'offres d'emploi et les pratiques des employeurs évoluent rapidement, l'analyse de l'utilisation des sites Internet par les recruteurs dans l'enquête Ofer de 2016 (Dares) fournit un cadrage du marché des types de sites d'offres d'emploi. Fin 2015, Pôle emploi était de loin le site le plus mobilisé : pour 62 % des recrutements ayant fait l'objet d'une annonce en ligne, une offre d'emploi a été diffusée sur le site de Pôle emploi (hors sites partenaires). Le site de l'entreprise était lui aussi largement utilisé par les recruteurs (40%) et devançait les *jobboards* généralistes (14 %) et spécialisés (18 %). Enfin, les recruteurs ont assez peu diffusé d'annonce en ligne sur les réseaux sociaux en 2016 (2 %). Si l'enquête Ofer n'a pas encore été renouvelée, on peut toutefois s'attendre à des changements conséquents dans le paysage des sites d'offres d'emploi, avec notamment une grande prise de part de marché des j*obboards* généralistes.

Choix des sites sources

En France, les recruteurs ont accès à de nombreux sites d'offres d'emploi en ligne. Pour cette première phase expérimentale, le nombre de sites a été limité (15 maximum) afin de ne pas démultiplier les problèmes techniques et administratifs liés à la collecte. Pour assurer une plus grande représentativité malgré leur nombre limité, les sites ont été choisis de manière à créer un ensemble couvrant tout le territoire, tous les types de métiers, de contrats et de qualifications. De plus, pour être retenus, les sites devaient diffuser des offres accessibles publiquement, sans avoir à se connecter à un compte utilisateur ou à payer par exemple. Pour assurer un volume minimum d'offres, des seuils ont été fixés : un stock journalier d'au moins 5 000 offres pour les sites spécifiques et 10 000 pour les sites généralistes. Enfin, les sites devaient être édités par des organismes dont le siège social est en France.
Après avoir expertisé plusieurs sites, un panel restreint mais diversifié a été élaboré. Il est composé principalement de *jobboards* français généralistes, d'un agrégateur et de sites plus spécialisés (postes de cadres, métiers du numérique…). La liste des sites scrapés a été (et sera encore) amenée à évoluer en fonction de la prise de part de marché de nouveaux sites, de la fermeture/fusion de certains sites déjà scrapés ou afin d'améliorer la couverture de certains métiers.

Récolte des données

Une fois les sources choisies, plusieurs possibilités existent pour collecter les offres. Trois méthodes de collecte automatisée d'offres d'emploi ont notamment été identifiées : le développement de partenariats avec des sites diffuseurs, la récupération de données *via* des API existantes et enfin le *webscraping (*ou *scraping*), qui consiste à collecter automatiquement des données depuis un site web. Cette dernière solution nécessite un investissement technique initial : le développement et la mise en place de scripts de *scraping* automatiques et spécifiques à chaque site. Ensuite, une maintenance régulière est nécessaire car il faut adapter les *scrapers* à chaque changement de structure des sites. Ainsi, plus le nombre de sites augmente, plus les coûts de maintenance sont élevés. Enfin, il existe un fort risque de blocage de la part des sites, qui peuvent souhaiter se protéger de ces techniques.
Mais la mise en place initiale peut être relativement rapide et l'accès aux données a l'avantage de se faire sans intermédiaire. Pour ces raisons, et car les coûts en maintenance sont relativement peu élevés pour le faible nombre de sites étudiés, la Dares a retenu cette solution pour tous les sites ne fournissant pas d'API (soit la quasi-totalité des sites hors Pôle Emploi).

Sélection et recodage des offres

Bien que les offres d'emploi répondent à une structure commune assez normée, la quantité d'informations qu'elles contiennent est variable selon les sites : certains proposent des offres très détaillées, alors que d'autres imposent un format plus concis. Ainsi, une offre doit contenir *a minima* les informations structurées suivantes pour être retenue : intitulé du poste, lieu de travail (code postal ou, à défaut, départemental) et nom de l'employeur. Ces variables ont été choisies car elles sont

disponibles sur la plupart des offres, correspondent aux champs nécessaires à l'étape de déduplication et sont utiles dans le cadre des analyses. En 2019, 9 % des offres scrapées sont retirées du champ car elles ne contiennent pas toutes les variables requises. Le taux d'offres inutilisables pour cette raison varie globalement de 0% à 10 % selon les sites.

Si les variables lieu de travail et nom de l'employeur sont utilisées telles quelles, un travail plus approfondi a été mené sur l'intitulé du métier. En effet, l'intégration des offres en ligne aux travaux de la Dares nécessite d'identifier le métier de l'offre. Plus, précisément, de classer correctement les offres dans une nomenclature officielle de métiers, le répertoire opérationnel des métiers et des emplois (Rome). Un algorithme de *machine learning* a été utilisé pour classer les offres par métier. Cet algorithme appelé *Support Vector Machine* (SVM) – est un algorithme d'apprentissage supervisé : il « apprend » à classer les offres à partir d'exemples d'offres d'emploi dont le métier est connu. Après avoir été entraîné sur des offres annotées en code Rome, il est capable de prédire le métier d'une offre avec un taux d'erreur d'environ 5% au niveau le plus fin. Les étapes nécessaires à l'entraînement de cet algorithme de *machine learning* (collecte des données annotées qui serviront d'exemples, nettoyage des données, entraînement de l'algorithme et évaluation des performances) sont détaillées dans le document d'études complet.

Déduplication des offres

Les offres d'emploi conservées dans la base finale sont dédupliquées, c'est-à-dire que si deux offres concernent vraisemblablement le même poste, seule une est conservée. Deux méthodes sont utilisées pour identifier les doublons : la déduplication par appariement des URL et la déduplication par proximité textuelle :

- La déduplication par appariement des URL consiste à identifier les URL présentes plusieurs fois dans la base et à n'en compter qu'une occurrence.

- La deuxième déduplication, par similarité textuelle, s'opère en deux étapes. La première étape consiste à regrouper les offres ayant le même code Rome, le même nom d'entreprise et le même lieu de travail (au niveau du code postal ou du code départemental, selon l'information disponible) qui ont été scrapées (« date_firstSeenDay ») à 14 jours de différence au maximum. À l'intérieur de chaque groupe, les offres sont comparées deux à deux : si leurs descriptions ont des vocabulaires semblables à plus de 95 % (ou plus exactement des vocabulaires ayant une similarité au sens de Jaccard supérieure à 0,95), elles sont considérées comme des doublons ; alternativement, elles sont supposées distinctes.

Afin que chaque offre de la base finale puisse être soumise à la déduplication par similarité textuelle, seules les offres comportant les informations nécessaires à la déduplication (code ROME, localisation et nom d'entreprise) sont conservées dans la base. La déduplication par proximité textuelle est calculée seulement sur les offres qui ne sont pas des doublons par url.

Représentativité des données

La relative facilité d'acquisition des offres en ligne et les informations nouvelles qu'elles délivrent contribuent à l'amélioration de la connaissance du marché du travail. Toutefois, les apports de cette source ne doivent pas en occulter les limites statistiques, et notamment les interrogations liées à sa représentativité. En l'absence de source – administrative ou d'enquête – de référence sur l'ensemble des offres ou sur les offres en ligne, l'usage statistique de Jocas ne peut reposer sur sa stricte représentativité, qu'il n'est pas possible de vérifier.

En effet, bien que l'enquête Ofer renseigne sur la part de marché de certains types de sites, il est difficile d'évaluer le champ couvert par un site ou un panel de sites. De plus, la source Internet est instable et mouvante : les émetteurs d'offres changent et se transforment rapidement et le champ couvert n'est donc *a priori* pas constant. Il est en particulier difficile d'évaluer la prise ou la perte de part de marché de certains sites. La représentativité des offres en ligne parmi l'ensemble des recrutements questionne également : les 49 % de recrutements faisant l'objet d'une annonce sur Internet (BERGEAT

et *al.*, 2018) ne sont pas représentatifs de l'ensemble des recrutements, le recours au canal Internet variant fortement selon les métiers. Enfin, certains phénomènes liés aux stratégies de recrutement (ex : offre unique pour plusieurs postes disponibles) compliquent encore l'interprétation des données.
Il n'en reste pas moins que Jocas est source de nouvelles informations et notamment d'indicateurs sur le fonctionnement du marché du travail en ligne.

Le document d'étude complet fournit des premiers cadrages, en comparant Jocas avec les sources usuelles exploitées pour effectuer le suivi du marché du travail, sur la base de leur couverture des métiers et du territoire.