AGOHRA : génération d'une ontologie dans le domaine des ressources humaines

Rémy Kessler[*] [1] — **Guy Lapalme**[**]

[*] *IRISA - UMR 6074, Université de Bretagne-Sud, 56017 Vannes, France*
kessler@univ-ubs.fr
[**] *RALI - Département d'informatique et de recherche opérationnelle*
Université de Montréal
C.P. 6128, Succ Centre-Ville, Montréal, Québec, Canada H3C 3J7
lapalme@iro.umontreal.ca

RÉSUMÉ. *Nous présentons une méthode d'analyse de corpus afin de générer une ontologie dans le domaine du e-recrutement. Notre approche de construction semi-automatique s'appuie sur des millions de profils issus de plusieurs réseaux sociaux et sur des dizaines de milliers d'offres d'emploi collectées sur Internet pour faire émerger des compétences et des connaissances communes afin de construire un ensemble structuré de termes et concepts représentant chaque métier. Notre approche combinant statistiques et extraction de n-grammes de mots permet de créer une ontologie en anglais et en français contenant 440 métiers issus de 27 domaines d'activité. Chaque métier est ainsi relié aux compétences nécessaires à sa pratique pour un total d'environ 6 000 compétences différentes. Une évaluation manuelle sur une portion de l'ontologie a été réalisée par un expert du recrutement et a montré des résultats de très bonne qualité.*

ABSTRACT. *We describe a corpus analysis method for generating an ontology in the field of e-recruitment from millions of user profiles gathered on social networks and tens of thousands of job offers collected over the internet. Using statistics and n-gram analyses, we create a structured set of terms and concepts in English and French for 440 occupations in 27 fields of activity. Each occupation is linked with the necessary practice skills for around 6000 different skills. A manual evaluation of the results was performed by a domain expert and has shown excellent results.*

MOTS-CLÉS : *ontologie, réseaux sociaux, ressources humaines, application industrielle.*

KEYWORDS: *ontology, social networks, human resources, industrial application.*

1. Travail effectué lors d'un stage postdoctoral à l'Université de Montréal

1. Introduction

Les développements rapides du Web et des réseaux sociaux durant la dernière décennie ont considérablement modifié la dynamique de recherche d'emploi comme le décrivent Sivabalan *et al.* (2014) . Les informations professionnelles publiées par les utilisateurs dans leurs profils (formations, antécédents de travail, résumé de carrière, liens sociaux, etc.) peuvent être exploitées par les recruteurs pour identifier de nouveaux candidats ou pour obtenir des informations complémentaires à leur propos.

D'après une étude de RegionsJob[1] (2011), « 43 % des recruteurs avouent recourir à des recherches de type nom/prénom sur les candidats qui postulent chez eux et 8 % des recruteurs interrogés déclarent avoir écarté un candidat à cause de traces jugées négatives trouvées en ligne ». La plupart de ces recherches étant effectuées de façon rapide et manuelle, les informations recueillies sur un individu à un premier niveau de recherche sont peu structurées, disparates, incomplètes, redondantes, parfois obsolètes et peuvent être biaisées, voire trompeuses (p. ex. à cause des homonymes).

La plupart des systèmes d'appariement entre une offre d'emploi et un profil s'appuient sur une ou plusieurs ressources linguistiques, coûteuses en entretien et en mise à jour. En effet, les métiers d'aujourd'hui ne sont pas forcément les mêmes que ceux d'hier, et ne seront peut-être pas identiques aux métiers de demain. L'évolution de notre société entraîne l'apparition de nouveaux métiers et de nouvelles compétences ainsi que la disparition d'autres. Les résultats d'une étude menée par la Harvard Business Review, citée par le journal *Libération*[2], indiquent que plus de 31 % de nouveaux métiers apparaissent chaque année et que 60 % des emplois actuels disparaîtront au cours des deux prochaines décennies. Afin de pallier ce problème, nous souhaitons développer un système pour générer une ontologie de façon semi-automatique en s'appuyant sur des données collectées sur Internet. Nous supposons que l'exploitation de ces informations est une voie prometteuse pour constituer un référentiel commun, une représentation de chaque domaine avec des liens logiques reliant chaque métier aux compétences nécessaires à sa pratique. Même si chaque profil est différent dans le détail, nous faisons l'hypothèse qu'un regroupement d'informations issues des mêmes métiers fera émerger des relations communes pour construire un ensemble structuré de termes et de concepts représentant chaque domaine. Même si la structure de l'ontologie reste simple, la génération de ces ressources, tout en restant automatisée, permettra de créer une représentation de la connaissance de chaque domaine qui pourra être exploitée par la suite dans le cadre de l'appariement (p. ex. pour évaluer si un candidat a toutes les compétences pour un métier) ou au travers d'expansion de requêtes (p. ex. afin de trouver les candidats les plus compétents pour un poste) ou de la génération de texte (p. ex. pour suggérer à un candidat comment mettre en avant ses compétences les plus en adéquation pour un poste).

1. https://entreprise.regionsjob.com/enquetes/reseaux_sociaux/resultats_enquete_2.pdf
2. http://www.liberation.fr/evenements-libe/2016/05/10/l-intelligence-artificielle-au-service-de-l-emploi_1451670

Dans le cadre du projet de recherche Butterfly Predictive Project[3] (BPP), nous développons une plateforme pour améliorer l'appariement entre des candidats et des offres d'emploi. Nous faisons l'hypothèse que l'acquisition et l'exploitation des traces laissées par les individus sur les réseaux sociaux (LinkedIn, Viadeo, etc.), diffuses, plus ou moins accessibles et peu structurées sont une voie prometteuse pour les recruteurs afin de déterminer le positionnement professionnel des candidats et faciliter leur mise en correspondance avec des emplois à pourvoir.

Même si de grands efforts ont été déployés ces dernières années afin de développer des ressources linguistiques et aider les systèmes d'appariement de candidatures et d'offres d'emploi, ces derniers se heurtent à la difficulté de créer des ontologies ou des taxonomies spécifiques à chaque domaine et particulièrement à leur entretien et leur mise à jour. Nous présentons ici une approche de génération dynamique d'ontologies avec une possibilité de mise à jour continue au fur et à mesure que le Web évolue.

Une version préliminaire de ce travail a été présentée à CORIA 2016 (Kessler *et al.*, 2016). Cet article présente plusieurs extensions telles que la recherche de compétences transversales, le dictionnaire dynamique ou encore le regroupement des synonymes qui ont grandement amélioré tant la qualité que la quantité des informations récoltées. Dans la section suivante, nous présentons des travaux liés à notre étude. La section 3 présente les ressources utilisées tandis que la méthodologie est détaillée en section 4. La section 5 décrit l'ensemble du système, avant de présenter les résultats et leur évaluation dans la dernière section.

2. Travaux connexes

De nos jours, des centaines de milliers de candidats mettent en ligne leur profil, et les entreprises ou les établissements publient une quantité importante de postes recherchés. Analyser automatiquement cette quantité d'informations pour mettre en correspondance emplois et candidats est une tâche difficile. Comme le décrivent Yahiaoui *et al.* (2006), cet appariement repose, d'une part, sur la connaissance des individus et de leurs compétences et, d'autre part, sur la connaissance des métiers. De grands efforts ont ainsi été déployés ces dernières années afin de constituer des ressources linguistiques pour améliorer les systèmes d'appariement.

Lau et Sure (2002) développent une ontologie, en se fondant sur une étude de cas de la société Swiss Life, centrée sur le domaine des technologies de l'information. Ils précisent que celle-ci a finalement été construite manuellement même si des approches semi-automatiques avaient été tentées, mais les résultats ne permettaient pas d'obtenir une représentation claire et structurée des compétences.

Les premiers travaux dans l'appariement de candidatures et d'offres d'emploi à l'aide d'une ontologie ont été proposés par Colucci *et al.* (2003 ; 2007). Leur système,

3. `http://rali.iro.umontreal.ca/rali/?q=fr/butterfly-predictive-project`

IMPAKT[4](Colucci *et al.*, 2013) permet d'extraire des compétences de CV et repose sur des méthodes de formalisation des raisonnements. Le système propose un appariement en effectuant des correspondances partielles ou complètes des compétences entre les offres d'emploi et les candidatures. Zimmermann *et al.* (2016) extraient les informations des CV en combinant apprentissage machine et traitement de la langue pour déterminer les candidatures les plus pertinentes en fonction d'un emploi déterminé.

Desmontils *et al.* (2002) et Trichet *et al.* (2004) décrivent une méthode d'indexation sémantique de CV. Celle-ci exploite les caractéristiques dispositionnelles du document afin d'identifier chacune des parties et de les indexer en conséquence. Ils proposent une instanciation de l'ontologie en partant des données récoltées et en s'appuyant sur des ressources externes (base ROME[5] et CIGREF[6]). Même si l'approche semble intéressante, l'absence de résultat ne permet pas d'évaluer l'apport de cette indexation particulière. Kmail *et al.* (2015) proposent une approche similaire en construisant des réseaux sémantiques à partir des candidatures et des offres d'emploi et enrichissent ceux-ci à l'aide de ressources externes telles que WordNet et YAGO2 (Hoffart *et al.*, 2011). Mochol et Simperl (2006) décrivent l'importance d'une ontologie commune (HR ontology) ainsi qu'un guide pour mettre en place ce type d'application tandis que Trichet *et al.* (2004) et Yahiaoui *et al.* (2006) décrivent différentes approches pour l'annotation sémantique de document et la gestion des compétences à l'aide d'ontologie dans le cadre du e-recrutement.

Dans le cadre du projet Prolix, Trog *et al.* (2008) décrivent une ontologie de ressources humaines en s'appuyant sur le cas de British Telecom. Ils proposent une architecture en plusieurs niveaux en fonction des compétences, des interactions et du contexte. Gómez-Pérez *et al.* (2007) proposent, quant à eux, une annotation sémantique des documents (offres d'emploi et CV) afin de construire différentes ontologies. Développées en anglais et disponibles en ligne[7], ces ontologies décrivent des compétences et des formations spécifiques au domaine des technologies de l'information tandis que d'autres sont plus généralistes comme les classes de métiers, les permis de conduire ou encore les secteurs d'activité.

Roche et Kodratoff (2006) présentent une extraction de terminologie sur un corpus de CV. Leur approche extrait un certain nombre de collocations contenues dans les CV sur la base de patrons (tels que nom-nom, adjectif-nom, nom-préposition-nom, etc.) et les classe en fonction de leur pertinence en vue de la construction d'une ontologie spécialisée.

Partant du constat que les médias sociaux deviennent une source incontournable pour les recruteurs dans leur recherche de candidats, Tétreault *et al.* (2011) décrivent

4. Information Management and Processing with the Aid of Knowledge-based Technologies.
5. Répertoire organisationnel des métiers et emplois : `http://www.pole-emploi.fr/candidat/le-code-rome-et-les-fiches-metiers-@/article.jspz?id=60702`
6. Club informatique des grandes entreprises françaises.
7. `http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/99-hrmontology`

la maquette d'une plateforme de recrutement intégrant les technologies du Web sémantique ainsi que l'élaboration d'une ontologie à l'aide du *Web Ontology Language* (OWL) dédiée au domaine TI. L'approche présente les avantages d'une application de ce type et différents scénarios de recrutement.

Plus récemment, le Vrang *et al.* (2014) présentent l'ontologie ESCO[8], un projet européen multilingue de classification de compétences, de métiers et de certifications afin de créer une harmonisation européenne en matière de recrutement. Cependant, même si le modèle est d'excellente qualité, il n'en est actuellement qu'à la version 0.1, ce qui restreint les domaines et les métiers couverts ainsi que les types de relation. Une autre limite de cette approche est son caractère rigide et figé. En effet, l'évolution d'une telle ressource est complexe puisqu'elle nécessite des connaissances dans plusieurs domaines ainsi que dans de nombreuses langues, ce qui passe par une longue phase manuelle de saisie et de vérifications.

La version actuelle regroupe 4 761 métiers et 5 096 compétences en 24 langues différentes pour environ 250 000 termes différents (21 000 termes en anglais et 18 000 en français). Chaque métier ou compétence est défini dans chaque langue par un *label préféré* ainsi que par un ou plusieurs *labels alternatifs* pouvant être un synonyme, une féminisation, une forme abrégée ou une variation orthographique. Chaque métier est en outre relié aux compétences nécessaires à sa pratique et de la même façon chaque compétence est reliée aux différents métiers. La version actuelle n'a qu'un seul type de relation `is-related-to`, mais d'autres relations sont prévues dans la prochaine version de cette ontologie[9].

Ces travaux montrent que différentes approches de construction d'ontologies dans le domaine du recrutement ont été envisagées, que ce soit avec des méthodes statistiques, sémantiques ou encore à l'aide de ressources linguistiques complémentaires. Même s'il existe d'autres approches et domaines pour lesquels des méthodes de construction semi-automatique d'ontologies ont été développées, par exemple l'astronomie (Bendaoud *et al.*, 2007), les dépêches journalistiques (Sami *et al.*, 2014) ou encore le domaine médical (Osborne *et al.*, 2009), nous observons que les travaux dans le domaine de l'e-recrutement reposent sur des créations manuelles d'ontologies et non sur une analyse de corpus.

L'originalité de notre approche est d'exploiter les réseaux sociaux et plus généralement l'information récoltée sur Internet pour mettre à jour la nomenclature des métiers et des compétences afin de construire un ensemble structuré de termes et de concepts et c'est ce que nous explorons dans cet article.

8. European Skills Competences and Occupations `https://ec.europa.eu/esco/`
9. `https://ec.europa.eu/esco/portal/escopedia/ESCO_data_model`

3. Données et connaissances

Une ontologie est un modèle de connaissances constitué de concepts relatifs à un domaine ainsi que des relations entre ces concepts. Afin de construire ce modèle, nous nous appuyons sur diverses données et connaissances, issues d'Internet ou fournies par notre partenaire industriel Little Big Job (LBJ). Le projet BPP s'appuie sur une modélisation en 44 secteurs d'activité (univers) utilisée par LBJ, chacun regroupant des familles de métiers assez proches et un ou plusieurs univers connexes. Les univers connexes à secteur donné sont ceux pour lesquels une transition est possible pour un candidat envisageant un changement de secteur. Par exemple, un candidat issu de l'univers *banque, finance, capital risque, fonds privés* pourra plus facilement envisager une transition vers les domaines connexes *assurances, mutuelles, prévoyance* ou *immobilier* que vers *cosmétique*.

Nous avons combiné ces connaissances avec l'ontologie issue du modèle ESCO décrit plus loin en 4, ainsi qu'un dictionnaire des métiers issus de la *Classification nationale des professions*[10] (CNP) et du *Répertoire opérationnel des métiers et des emplois (ROME)*[11]. La CNP est la référence reconnue des professions au Canada tandis que le ROME est la référence en France. La CNP répartit plus de 40 000 appellations d'emploi (anglais/français) en 500 profils de groupes professionnels tandis que le ROME comprend 10 000 appellations différentes réparties en 531 groupes ; pour notre projet, notre partenaire industriel nous a suggéré de nous limiter à un sous-ensemble de domaines qui l'intéressent soit 3 730 métiers. Nous avons par ailleurs observé des appellations différentes entre ROME et CNP suivant le pays (instituteur/enseignant, commis de bureau/réceptionniste, garde forestier/forestier, etc.), ce qui nous a conduits à conserver les deux classifications.

Par ailleurs, plus de dix millions de profils issus de plusieurs réseaux sociaux (*LinkedIn*, *Viadeo*, *Indeed* et d'autres) ont été récoltés à l'aide d'un processus de collecte automatique de sites Internet. Ces données issues de profils publics professionnels ont été préalablement anonymisées puis agrégées dans un format uniforme. L'origine géographique étant le Canada ou la France, les profils sont soit en français soit en anglais ou encore bilingues. Chaque profil résume différentes informations sur le parcours du candidat telles que ses diplômes et ses formations ainsi que leurs dates. De la même façon, une section du profil rend compte de ses expériences. Chacune contient plusieurs éléments tels que les dates de début et de fin, le nom de la société employeur (avec éventuellement une URL vers sa page Internet), la fonction occupée par le candidat au cours de cette expérience ainsi que le lieu et un éventuel descriptif de sa mission au sein de cette société. Un résumé des expériences sous forme d'un texte est par ailleurs présent ainsi qu'une courte description sous la forme d'une phrase d'accroche permettant au candidat de se décrire en quelques mots. D'autres informations sont récoltées ou éventuellement calculées telles que l'expérience totale du candidat,

10. http://www5.hrsdc.gc.ca/noc/
11. http://www.pole-emploi.fr/candidat/le-code-rome-et-les-fiches-metiers-@/article.jspz?id=60702

```
{ ...
"countryCode": "FR",
"createdDatetime": "2012-12-19 19:56:47",
"city": "Vannes",
"personalBrandingClaim": "Chef de projets au CHU de Vannes",
"personalBrandingPitch": "Professionnelle dynamique et proactive ayant eu l'opportunité de
           développerdes compétences variées tant dans le milieu de la santé que dans celui de
           l'aéronautique . Principalement axée sur la gestion, l'amélioration continue
           des processus et des pratiques et l'atteinte des objectifs"},
"educations":[{"name": "Marketing pharmaceutique",
              "schoolName": "Universite Claude Bernard (Lyon I)"},
             {"name": "MBA ", "schoolName": "Universite de Montreal - HEC Montreal"}],
"experiences":[{"function": "Chef de projets",  "companyName": "CHU de Vannes"},
              {"function": "Directrice du service 'a la clientele pharmaceutique",
               "companyName": "zootopia"},
              {"function": "Coordonnatrice de projets",  "companyName": "chu-saint-Trudeau"}]
"skills": ["E-commerce", "SDL Fredhopper", "Gestion de projet", "Gestion d'équipe",
          "Accessibilité"],
"languages": [{language:"Français", "level": "Native or bilingual"},
             {language:"Anglais", "level": "Native or bilingual"}],
...
}
```


Figure 1. *Extrait de la structure JSON d'un profil créé à partir d'informations collectées sur des réseaux sociaux*

	Canada	France
Nombre de profils	2 658 467	7 484 311
Moyenne du nombre d'expériences	3,2	2,3
Moyenne du nombre de formations	1,39	1,06
Moyenne du nombre de compétences	0,17	0,12
Statistiques textuelles des profils		
Profil vide	9,61 %	16,34 %
Moins de 100 caractères	26,95 %	40,53 %
Moins de 300 caractères	16,18 %	12,70 %
Moins de 500 caractères	6,43 %	4,66 %
Plus de 500 caractères	40,83 %	25,77 %

Tableau 1. *Statistiques de la collection de profils de réseaux sociaux*

les langues qu'il maîtrise, ses loisirs, le nombre de relations avec d'autres candidats ou encore les compétences acquises au cours de son parcours professionnel. La figure 1 présente un exemple de profil avec des informations extraites de réseaux sociaux.

Chaque profil regroupe une cinquantaine de champs, mais il existe cependant un nombre important de profils ne contenant pas ou peu d'informations comme le montre le tableau 1 qui présente quelques statistiques descriptives de cette collection. Dans le cadre de cette application, nous nous concentrons sur les 2,3 millions de compétences issues de ces profils.

En complément de ces données, 300 000 offres d'emploi ont été collectées sur Internet, 200 000 en anglais et 100 000 en français. Ces offres d'emploi couvrent un grand nombre de métiers et sont issues, elles aussi, du Canada ou de la France. Chaque offre d'emploi contient un titre, une description contenant le détail de l'offre d'emploi, la date de mise en ligne, le lieu de la mission proposée ainsi que le nom de la compagnie qui recrute. Les premières observations ont montré que, bien qu'extrêmement bruitées, ces données peuvent constituer une source intéressante d'informations afin de constituer une ontologie de façon semi-automatique.

4. Méthodologie

Nous présentons tout d'abord le modèle ESCO (le Vrang *et al.*, 2014) et sur lequel s'appuie notre démarche. Celui-ci est organisé selon le schéma de modélisation SKOS (Miles et Bechhofer, 2009). Il classe les connaissances disponibles en trois piliers : professions, compétences et qualifications. Dans le cadre de ces travaux, nous nous concentrons sur les professions et les compétences. Le modèle se structure par la suite en concepts et en termes. Chaque profession, compétence et qualification est associée à un concept et est identifiée de façon unique par un *uniform resource identifier (URI)*. Chaque concept comporte au moins une étiquette préférée alors que des possibles synonymes, des variantes d'orthographe et des abréviations sont enregistrés comme des étiquettes alternatives. Des relations sont créées manuellement entre chaque profession et les compétences nécessaires à leur pratique.

Dans un premier temps, nous avons effectué une comparaison entre les compétences issues des réseaux sociaux et celles issues de l'ontologie ESCO. Les résultats ont montré de nombreuses correspondances exactes ou partielles (plus d'un million) et variées. Les correspondances partielles ont été calculées, dans un premier temps, en comparant uniquement les quatre premiers caractères de chaque compétence puis dans un second temps avec la mesure de Levenshtein (Levenshtein, 1966). Nous observons que les compétences les plus fréquentes dans les données issues des réseaux sociaux se retrouvent bien dans l'ontologie ESCO.

Cette approche a cependant montré certaines limites. Même si le modèle ESCO est d'excellente qualité, les domaines et métiers couverts par l'ontologie restent limités (par exemple, on ne retrouve pas *business analyst* ou *account manager*, pour les métiers, ni *marketing strategy*, *financial modelling*, *food cost management* pour les compétences). Un enrichissement d'ESCO a été envisagé, cependant le modèle de 44 univers du partenaire industriel était incompatible avec les regroupements par domaines effectués dans ESCO. Quant aux classifications ROME et CNP (décrites dans la section précédente), elles ont été développées dans un cadre administratif afin de codifier les professions pour les besoins des bureaux d'immigration et de statistiques gouvernementales. Si ces organismes cherchent bien à couvrir tous les métiers (y compris celui de Premier ministre ou député), ils ont recours à une nomenclature très différente : nous n'observons que 73 appellations identiques entre les classifications ROME et CNP, 29 entre ESCO et ROME et seulement 13 entre ROME, CNP

et ESCO ; de plus, ces classifications gouvernementales ne font aucun lien entre métiers et compétences qui est pourtant le but que nous cherchons à atteindre avec notre ontologie.

Par ailleurs, il est extrêmement difficile de discerner dans les profils des candidats les compétences issues d'une expérience plutôt que d'une autre : par exemple, le profil d'un chef de projet senior en technologies de l'information cumulera des compétences en tant que développeur issues de ses premières expériences, avec celles, par la suite, de chef de projet. Il est donc difficile de se restreindre à cette unique source pour déterminer les compétences requises pour un emploi.

Afin de pallier ces problèmes, nous avons décidé d'utiliser comme source de documents pour l'ensemble des offres d'emploi. En effet, ces offres définissent les expériences et qualifications désirées pour un métier donné. Comme nous disposons d'un nombre important d'offres d'emploi (300 000), nous croyons être en mesure de faire émerger les compétences et les connaissances communes pour construire un ensemble structuré des termes et des concepts représentant chaque métier. L'ensemble des offres d'emploi étant collecté sur Internet, le modèle pourra s'enrichir, à terme, de façon semi-automatique avec l'apparition et la disparition de métiers dans les offres d'emploi.

5. Vue d'ensemble du système AGOHRA

La figure 2 présente une vue d'ensemble du système AGOHRA[12] dont les étapes seront détaillées dans le reste de la section. Au cours d'une première étape ①, nous effectuons une normalisation des offres d'emploi. Le module suivant ② utilise les titres des offres d'emploi afin de détecter les métiers avant d'y associer un univers ③ (voir section 3 pour la notion d'univers). Une première recherche au cours de l'étape ④ est effectuée par la suite, afin de constituer un dictionnaire composé uniquement de *compétences transversales*[13]. L'étape suivante ⑤ consiste à utiliser l'ensemble du vocabulaire récolté afin de faire émerger les compétences. À l'aide de la base de profils ainsi que de dictionnaires constitués de façon dynamique, le module suivant ⑥ classe par la suite le vocabulaire obtenu afin de déterminer s'il s'agit de compétences. Le dernier module ⑦ transforme ensuite les informations ordonnées et structurées en une ontologie au format RDF.

5.1. *Normalisation des offres d'emploi*

Nous effectuons au préalable une extraction du contenu textuel des offres d'emploi au format HTML. Au cours de l'étape ①, nous effectuons une séparation des

12. *Automatic Generation of an Ontology for Human Resource Applications*
13. Les compétences transversales, aussi appelées *soft skills*, sont des compétences personnelles et sociales, orientées vers les interactions humaines.

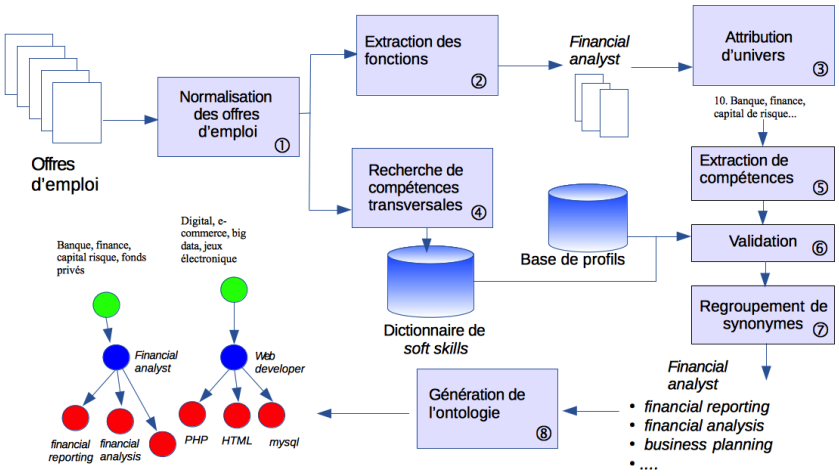

Figure 2. *Vue d'ensemble du système AGOHRA*

offres en français et en anglais ainsi qu'un filtrage des quelques offres dans les autres langues (environ 150 offres sont en espagnol, japonais, italien, etc.). L'observation de l'ensemble des offres montrant un grand nombre de doublons (offres similaires avec des dates différentes, offres similaires pour des lieux différents, etc.), nous effectuons un dédoublonnage des offres d'emploi afin d'éviter d'augmenter artificiellement les fréquences des termes présents dans plusieurs offres pratiquement identiques. Environ 6 000 offres d'emploi sont ainsi écartées. Toujours au cours de cette étape, une normalisation des titres des offres d'emploi est effectuée afin de détecter les métiers en utilisant différents patrons afin d'améliorer la qualité du processus. On supprime ainsi les accents, les caractères particuliers tels que « - », « / », « ( ) » ou encore les expressions couramment utilisées dans les titres d'offres d'emploi telles que *full-time*, *temporary*, ou encore *entry level*. Différents processus linguistiques sont utilisés afin de réduire le bruit dans le modèle : les expressions courantes (par exemple, *c'est-à-dire*, *chacun de*, etc.), les chiffres et les nombres (numériques et/ou textuels), les symboles spéciaux ainsi que les termes contenus dans un antidictionnaire adapté à notre problème.

Nous avons également constitué une liste de noms de métiers normalisés regroupant sous un terme unique les différentes écritures pour chaque métier (féminin, pluriel, erreurs typographiques, etc.) classées selon leur fréquence d'apparition dans le

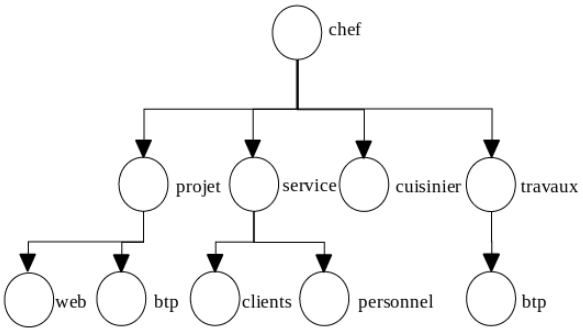

Figure 3. *Extrait de l'arbre de métiers avec* chef *comme premier nœud ce qui permet de retrouver* chef projet web*, ...* chef cuisinier*, ...* chef travaux btp*.*

champ, en fonction des expériences issues des profils de réseaux sociaux. On observe par exemple 292 façons différentes (bien comptées !) d'écrire la fonction *développeur* dans l'ensemble (p. ex. *développeur, developper, développeuse, dévelloppeur, developer, developppper,* etc.). Un processus de lemmatisation avait été intégré lors des premières expériences, mais il s'est avéré peu performant, car la plupart des noms d'emploi ne sont pas déclinés et ne posaient aucun problème compte tenu de cette normalisation qui est mieux adaptée à notre application.

5.2. *Extraction des fonctions et attribution d'univers*

Nous utilisons au cours de cette étape ② les titres des offres d'emploi normalisés (voir 5.1) afin de construire une liste de métiers. À l'aide de règles, le système compare les titres avec un arbre préfixe contenant la liste de métiers normalisés décrite dans la section 5.1. Chaque métier est ainsi transformé en un ensemble de nœuds où il existe un nœud pour chaque mot, comme le montre l'exemple de la figure 3.

Cette structure permet d'effectuer des comparaisons rapides entre la liste des métiers et les titres tout en conservant les variations d'écriture. Nous conservons ainsi la chaîne la plus longue comme métier reconnu. Grâce à un seuil de fréquences minimal déterminé empiriquement, nous ne conservons que les métiers avec une fréquence importante. Une première version du système traitait de l'ensemble des métiers présents dans les offres d'emploi, mais nous avons décidé de privilégier les métiers ciblés par LBJ, notre partenaire industriel, dont l'activité est centrée sur le recrutement dans les domaines de la gestion, de la finance ou de la haute technologie.

En retirant les métiers qui n'intéressaient pas notre partenaire industriel, comme *gardienne de chat*, *nounou*, *camionneur*... nous avons filtré une grande quantité d'offres d'emploi. Nous en conservons tout de même 45 000 (25 000 en anglais et 20 000 en français). L'étape ③ consiste à attribuer un univers à chacun des noms de métiers obtenu.

Afin d'associer un univers à chaque nom de métier, nous avons expérimenté plusieurs approches à l'aide d'algorithmes de type machine à vecteurs de support ou de *boosting*. Nous avons transformé le texte en vecteurs de mots afin de détecter de manière automatique l'univers associé à chaque métier. Après avoir découpé l'ensemble des offres d'emploi en cinq sous-ensembles approximativement de la même taille, nous avons appliqué la procédure suivante : quatre des cinq sous-ensembles ont été concaténés pour produire un corpus d'entraînement et le cinquième a été utilisé pour le test. La procédure a été effectuée cinq fois afin que chacun des sous-ensembles du corpus d'apprentissage soit utilisé une fois pour le test. Cependant, les chevauchements de vocabulaire entre univers pour certaines offres d'emploi ont rendu la tâche délicate (par exemple des offres d'emploi d'analystes financiers travaillant pour des entreprises qui ne sont pas dans le domaine de la finance ou des postes en informatique pour un groupe de cosmétique, etc.). Nous avons finalement décidé d'associer manuellement à cette liste un univers pour chaque métier.

5.3. *Extraction de compétences*

L'objectif de l'étape ⑤ est d'utiliser le vocabulaire issu des offres d'emploi afin de faire émerger les compétences associées à chacun des métiers. Nous effectuons pour cela une agrégation du vocabulaire en le regroupant par métier. L'extraction d'information dans une offre d'emploi n'est pas une tâche triviale comme le soulignent Loth *et al.* (2010). Kessler *et al.* (2008) montrent qu'en raison d'une grande variété dans les paramètres (texte libre, tailles différentes, découpage incertain, délimiteurs variés), le découpage d'offres d'emploi en blocs d'information est une tâche délicate. Ces offres apparaissent cependant dans un ordre conventionnel. Afin de diminuer la taille du vocabulaire considéré, nous recherchons différents motifs séparateurs et fréquents dans une offre d'emploi, tels qu'*exigences*, *qualifications*, *responsibilities*, etc. Ces motifs, bien que pas toujours présents, permettent ainsi de réduire considérablement le vocabulaire en ne prenant en compte que la partie de l'annonce suivant le motif.

L'observation des compétences de l'ensemble des profils de réseaux sociaux décrit en section 3 (champs `skills` de la figure 1) montre que plus de 80 % des compétences se présentent sous la forme de n-grammes de mots (par exemple *financial modelling*, *php development*) répartis comme suit : 24 % d'unigrammes, 42 % de bigrammes et 20 % de trigrammes, 8 % de 4-grammes, 4 % de 5-grammes et 1 % vide (c'est-à-dire des profils sans aucune compétence). Compte tenu de ces observations, nous avons décidé de transformer l'ensemble du vocabulaire issu des offres d'emploi sous forme d'unigrammes, de bigrammes et trigrammes et de les ordonner selon un score S_{job}, inspiré du *TF-IDF* :

$$S_{job}(u, m) = tf(u).log\frac{D_m}{df(u)} \quad [1]$$

avec u l'unité lexicale considérée (unigrammes, bigrammes ou trigrammes), $tf(u)$ la fréquence de u dans la collection, $df(u)$ le nombre de documents où l'unité lexicale u apparaît et D_m le nombre de documents associés au métier m.

Afin de filtrer certaines compétences courantes dans les offres d'emploi telles que les compétences « Microsoft » (par exemple *microsoft office*, *suite office*, *word*, etc.) et qui viennent parfois bruiter les résultats, nous avons mis en place un certain nombre de règles complémentaires. Nous avons fait de même pour les compétences de langues (*anglais écrit*, *français*, *bilingue*, etc.) qui, bien qu'essentielles, n'étaient pas des compétences représentatives de chaque métier. Ces règles ont été développées pour répondre à des suggestions de LBJ, notre partenaire industriel, suite à l'examen des résultats d'une version préliminaire du système.

5.4. *Recherche de compétences transversales*

Les premiers résultats ont montré qu'un grand nombre de compétences extraites étaient des *compétences transversales* (telles que *verbal/written communication skills*, *capacité à travailler en équipe*, etc.) qui peuvent être considérées comme pertinentes quel que soit le métier considéré. Nous avons donc ajouté l'étape ④ afin de les distinguer des compétences plus techniques, communément appelées *hard skills*[14]. Les compétences transversales étant des compétences demandées dans l'ensemble des offres d'emploi et quelle que soit la fonction considérée, nous effectuons un premier traitement en réutilisant l'ensemble des offres d'emploi sans tenir compte des fonctions. Nous recherchons ainsi les compétences les plus fréquentes quel que soit le métier considéré. Après l'étape de normalisation (section 5.1), nous effectuons une agrégation du vocabulaire décrite en section 5.3, que nous transformons par la suite sous forme d'unigrammes, de bigrammes et de trigrammes et que nous ordonnons en fonction de S_{job}, tel que défini par l'équation [1]. L'ensemble des compétences obtenues permet de constituer un dictionnaire qui sera utilisé par la suite, au cours de l'étape de validation (section 5.5), afin de séparer les compétences transversales des compétences plus techniques. Nous obtenons ainsi un dictionnaire de 250 termes en français et en anglais, regroupant unigrammes, bigrammes et trigrammes.

5.5. *Validation des compétences*

Une première version du module ⑥ effectuait une comparaison de la liste ordonnée de n-grammes obtenus avec un dictionnaire contenant environ 25 000 compétences issues des profils de réseaux sociaux afin de valider ou invalider les n-grammes. Cette méthode était relativement efficace pour écarter les n-grammes qui n'étaient pas des compétences tels certains termes ou expressions courants dans les offres d'emploi (par

14. Les *hard skills* sont les compétences formellement démontrables, nées d'un apprentissage technique, souvent d'ordre académique, et dont la preuve est apportée par l'obtention de notes, de diplômes, de certificats.

exemple *employment equity*, *strong experience required*, etc.). Mais elle ne permettait pas de prendre en considération la spécificité de certaines compétences vis-à-vis de certains métiers. Afin de résoudre ce problème, nous avons constitué un dictionnaire *dynamique*. Pour le construire, les 10 millions de profils ont été indexés avec le moteur de recherche Lucene[15] en fonction de la langue du profil et du pays d'origine. Chaque champ textuel a été indexé individuellement afin de pouvoir être interrogé séparément. Cette approche affine les résultats en fonction du secteur d'activité, de la fonction occupée par le profil ou des compétences qu'il possède. Pour chaque métier sélectionné au cours de l'étape 5.2, nous effectuons une requête avec le nom de ce métier et en spécifiant que les résultats doivent contenir uniquement des profils dont la section *compétences* n'est pas vide. Les profils retournés sont donc des profils qui occupent ou qui ont occupé le métier recherché et dont la section *compétences* a été renseignée. Pour une requête donnée, nous agrégeons les compétences des 10 000 premiers profils[16] obtenus. Les compétences sont ainsi transformées en une liste ordonnée par fréquence avec comme seuil minimal 10 % de la fréquence maximale obtenue. Chaque liste constitue ainsi un dictionnaire *dynamique* des compétences les plus fréquentes pour chaque métier selon la base de profils de réseaux sociaux. Nous comparons la liste ordonnée de n-grammes obtenus avec ce dictionnaire dynamique et avec le dictionnaire de *soft skills* afin de séparer les compétences en fonction de leur type comme expliqué en section 5.4.

5.6. *Regroupement de synonymes*

L'analyse de la première version de l'ontologie a montré qu'un grand nombre de compétences, seul l'adjectif étant différent, pouvaient être regroupées en une seule (telles que *good work ethic / strong work ethic*, *bonne capacité adaptation / excellente capacité adaptation*, etc.). Un module spécifique ⑦ recense l'ensemble des adjectifs issus de la collection de profils de réseaux sociaux afin de constituer une liste des plus fréquents. Même si certains d'entre eux sont classiques (par exemple *good*, *excellent*), d'autres n'ont de sens que dans le domaine particulier du recrutement et de la qualification de compétences (tels que *strong*, *outstanding*). Nous utilisons par la suite cette liste afin de constituer plusieurs motifs pour sélectionner, pour chaque type de n-grammes, un sous-ensemble de compétences susceptibles d'être des synonymes une fois l'adjectif retiré. Toutes les occurrences des synonymes sont ensuite regroupées sous une compétence unique, les adjectifs classés par ordre alphabétique et séparés par un « / » (par exemple, *excellent/good/strong communication skills*).

15. http ://lucene.apache.org

16. Ce nombre a été déterminé de façon empirique.

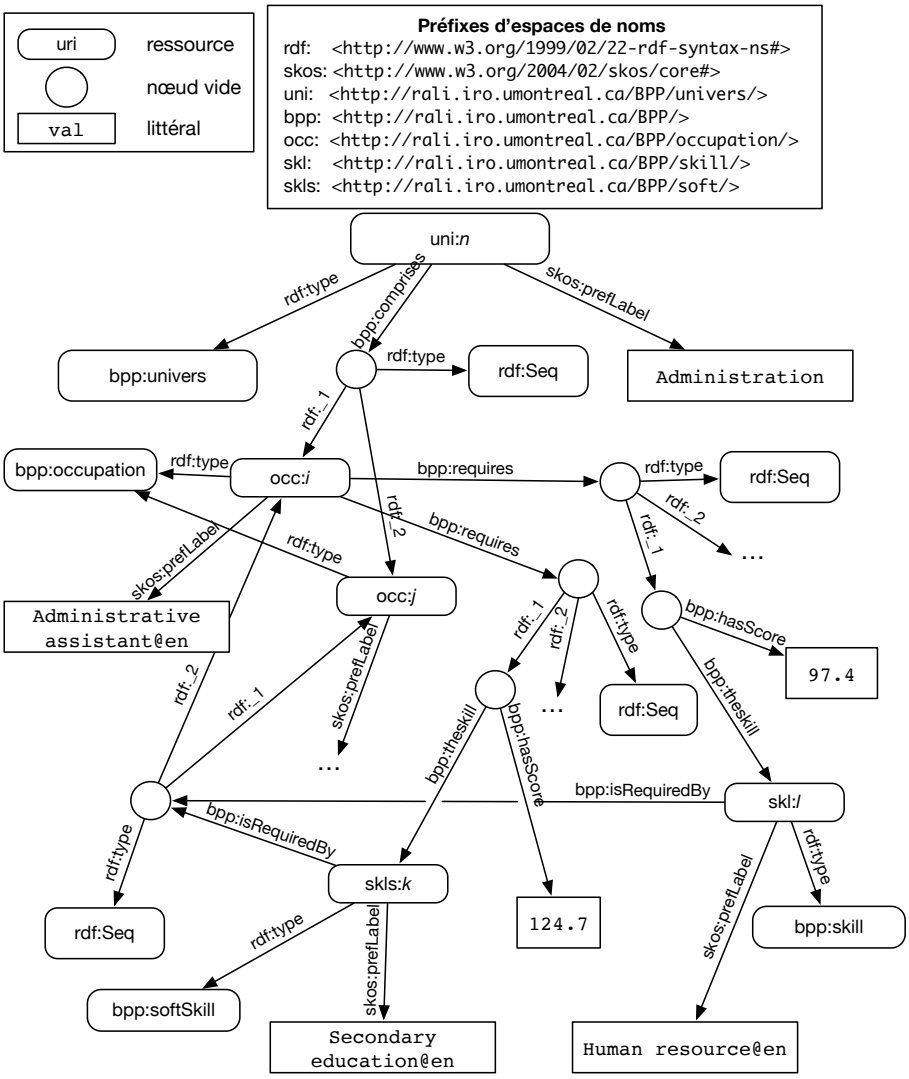

Figure 4. *Organisation du graphe RDF de l'ontologie générée*

5.7. *Création de l'ontologie*

À la suite de l'application des traitements linguistiques précédents sur l'ensemble de nos sources de données (bases de données spécialisées ou textes glanés du Web), nous avons décidé d'organiser le résultat sous forme d'une ontologie dont la structure reprend des éléments de l'ontologie ESCO (liens entre métiers et compétences), mais en y intégrant la notion d'univers définie par LBJ (section 3).

Notre ontologie est une classification hiérarchique, exprimée en RDF, dans laquelle chaque univers (`uni:n` dans la figure 4) est lié par la relation `bpp:comprises` à un ensemble de métiers `occ:i` nommés *occupations* par souci de compatibilité avec la nomenclature ESCO. Chaque *occupation* est liée par la relation `bpp:requires` à deux séquences (`rdf:Seq`) de compétences : une pour les compétences transversales (`bpp:softskill`) et une autre pour les compétences spécialisées (`bpp:skill`). Chaque compétence est liée à une étiquette par la relation `skos:prefLabel` sous forme de n-grammes et associée à un score `bpp:hasScore` qui est la valeur de S_{job}. Chaque métier est ainsi rattaché à son domaine ainsi qu'à une liste des compétences associées. De plus, un lien direct `bpp:isRequiredBy` est ajouté entre chaque compétence et le métier qui l'exige. Même si chaque compétence est unique, celle-ci peut être reliée à plusieurs métiers. Notons que conformément aux principes du Web sémantique, une classe RDF définit l'ensemble des individus qui partagent des propriétés communes, il n'y a donc pas d'instanciation d'objets à partir des classes comme dans les modèles à objets plus *classiques*.

Le module ⑧ crée une ontologie[17] composée des titres de 440 métiers (128 en anglais et 312 en français) répartis dans 27 univers différents (sur les 44 univers possibles) et reliés à 6 226 compétences différentes (4 059 pour l'anglais et 2 167 pour le français) et 485 compétences transversales (259 pour l'anglais et 226 pour le français). Des exemples de deux métiers avec leurs compétences identifiées sont présentés dans les tableaux 3 et 4. Plusieurs métiers peuvent partager des compétences qui ne sont pas répétées dans l'ontologie. La structure permet également de trouver tous les métiers qui demandent une compétence particulière.

Le réseau RDF pour l'anglais contient 109 156 triplets pour 128 métiers reliés à 4 059 compétences spécifiques et 259 compétences transversales et celui pour le français comprend 160 584 triplets pour 312 métiers reliés à 2167 compétences spécifiques et 226 compétences transversales. Ils sont interrogeables avec SPARQL dans nos applications. Pour en faciliter l'exploration visuelle, nous avons aussi développé un navigateur spécialisé qui affiche la structure RDF en faisant ressortir les différents niveaux : univers, métiers et compétences. La figure 5 présente une capture d'écran du navigateur contenant un extrait de l'ontologie finale avec un des métiers évalués en section 6. Le champ de texte en haut du navigateur permet de rechercher des métiers ou des occupations qui contiennent une chaîne. Il est ensuite possible d'explorer les compétences ou les métiers associés. Ce navigateur regroupe les résultats sous forme de triplets pour en faciliter l'évaluation, mais la structure de l'ontologie ne repose pas sur cette distinction qu'il est simple d'ignorer lors de requêtes SPARQL.

Cette première ontologie a une structure relativement simple, mais suffisante pour les besoins de notre application : elle est composée de listes d'occupations regroupées en univers, les occupations étant reliées à des listes de termes identifiant les compétences requises par ces occupations. Il aurait été intéressant de hiérarchiser les métiers

17. Disponible à `http://www-labs.iro.umontreal.ca/~lapalme/LBJ/BPPontologie/`

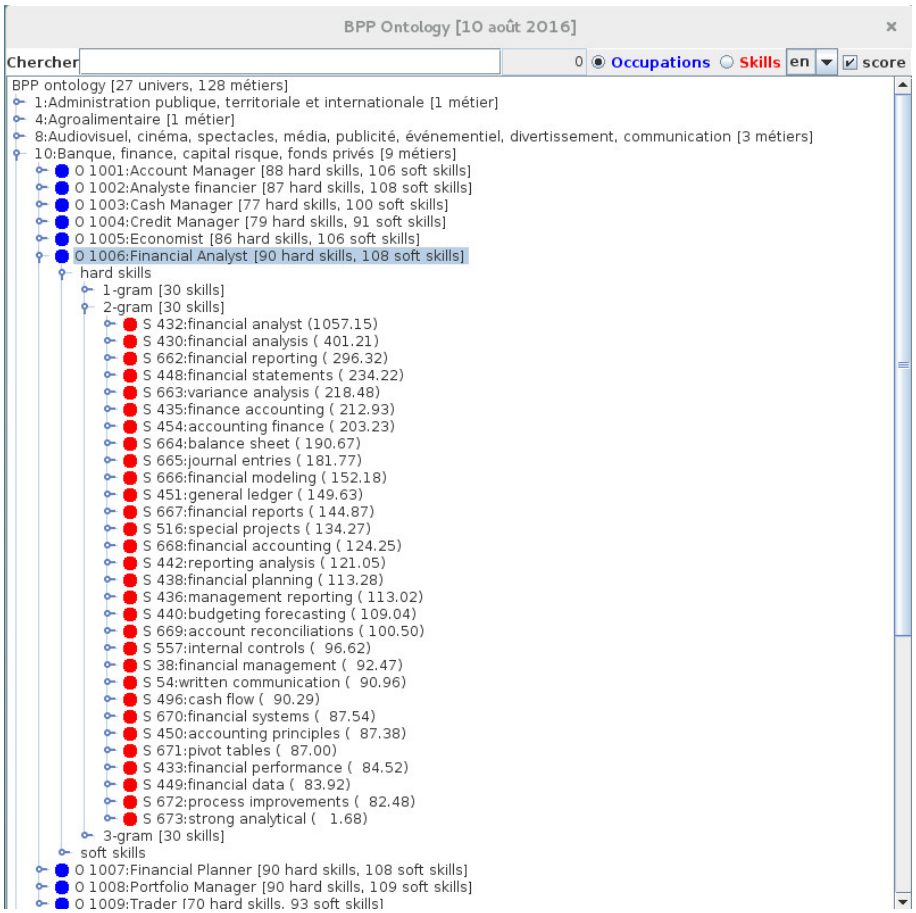

Figure 5. *Navigateur pour rechercher dans l'ontologie. Chaque concept présent dans l'ontologie est déterminé à l'aide d'un identifiant numérique, les compétences sont précédées de la lettre S (*Skills*) les métiers par O (*Occupations*), suivi du numéro associé à l'univers (par exemple O1006* financial analyst *appartient à l'univers 10* banque, finance, capital risque, fonds privés*). Le nombre entre parenthèses est la valeur du score calculée selon l'équation* [1] *(section 5.3).*

(p. ex. regrouper les *chefs de projets* de tous les domaines) ou de créer une structure d'héritage entre les compétences, mais ceci fera l'objet d'un travail futur.

6. Évaluation

Nous présentons maintenant nos résultats ainsi que son évaluation. Les tableaux 3 et 4 présentent les compétences obtenues, sous forme de n-grammes classés arbitrairement en fonction de la taille. Nous ne présentons ici que les dix premières compétences en ordre décroissant de S_{job} (voir section 5.3) pour les métiers *analyste financier* (Tableau 3) et *analyste programmeur* (Tableau 4) en français et en anglais. L'évaluation de cette portion de l'ontologie a cependant été effectuée sur les trente premières compétences par un expert du domaine. Nous avons choisi ces métiers afin de privilégier des métiers ciblés par notre partenaire industriel et pour ne pas nous limiter aux métiers des technologies de l'information, habituellement utilisés pour les évaluations dans la littérature. Nous présentons dans le tableau 2 un aperçu des métiers contenus dans l'ontologie dans chacune des langues, classés en fonction du nombre d'offres d'emploi utilisées par le système.

Anglais		Français	
Métier	# offres	Métier	# offres
sales representative	2 211	commercial	3 159
designer	1 764	développeur	2 692
administrative assistant	1 644	technicien commercial	1 451
customer service representative	1 309	comptable	1 363
account manager	1 183	contrôleur de gestion	1 000
developer	1 178	ingénieur commercial	983
business analyst	1 053	responsable commercial	732
store manager	918	acheteur	706
truck drive	729	ingénieur études et développement	570
sales manager	665	responsable comptable	479
assistant manager	621	responsable ressource humaines	422
restaurateur	565	chef produit	407
financial analyst	497	architecte	380
business developer	496	responsable production	373
account executive	454	infirmier	340

Tableau 2. *Liste des métiers dans chacune des langues, classés par ordre décroissant du nombre d'offres d'emploi. Seuls les 15 premiers sont présentés ici, mais les résultats génèrent 128 métiers en anglais et 312 en français.*

Financial analyst	
Soft skills	Hard skills
financial, business, support, management, process, reports, data, project, ~~including~~, projects ...	accounting, analysis, finance, reporting, cpa, cma, budget, cga, ~~end~~, forecast ...
analytical skills, communication skills, problem solving, ability work, internal external, ~~real estate~~, decision making, interpersonal skills, financial services, verbal written ...	~~financial analyst~~, financial analysis, financial reporting, financial statements, variance analysis, finance accounting, accounting finance, balance sheet, journal entries, financial modelling ...
analytical problem solving, problem solving skills, verbal written communication, ability work independently, key performance indicators, fast paced environment, oral written communication, time management skills, communication interpersonal skills, interpersonal communication skills ...	financial planning analysis, ad hoc reporting, financial reporting analysis, financial analysis reporting, financial statement preparation, year end close, consolidated financial statements, planning budgeting forecasting, business case analysis, possess strong analytical ...

Analyste financier	
Soft skills	Hard skills
comptabilite, analyse, connaissance, information, gestion, direction, ~~recherche~~, environnement, organisation, experience	cpa, finance, finances, consolidation, qualifications, cma, cga,~~principal~~, ~~ecrit~~, bilinguisme ...
etats financiers, experience client, esprit analyse, capacite travailler, analyse synthese, travail equipe, relations interpersonnelles, administration affaires, resolution problemes, gestion priorites ...	analyse financiere, analyses financieres, amelioration processus, processus budgetaire, modelisation financiere, processus affaires, financial reporting, financial analyst, cycle comptable, prix revient ...
cycle comptable complet, capacite travailler pression, esprit analyse synthese, capacite analyse synthese, facilite travailler equipe, capacite travailler equipe, ~~word power point~~, problem solving skills, communication orale ecrite, strong analytical skills ...	consolidated financial statements, analysis problem solving, ~~strong business acumen~~, excellent organizational skills

Tableau 3. *Liste de compétences classées par unigrammes, bigrammes, trigrammes pour les métiers* financial analyst *et* analyste financier *obtenue à partir de respectivement 497 et 127 offres. Les n-grammes considérés comme non pertinents ont été barrés. Les compétences sont triées en ordre décroissant de score.*

Programmer analyst	
Soft skills	Hard skills
~~development~~, technical, systems, software, application, business, design, support,~~data~~, solutions	~~programmer~~, ~~analyst~~, programming, applications, sql, java, test, web, developing, integration
computer science, software development, information technology, business requirements, working knowledge, problem solving, internal external, best practices, ~~experience working~~, communication skills	application development, sql server, design development, web services, unit testing, asp net, java developer, web applications, production support, vb net
problem solving skills, analytical problem solving, ability work independently, verbal written communication, written verbal communication, subject matter expert, written oral communication, build strong relationships, verbal communication skills, communication interpersonal skills	object oriented programming, service oriented architecture, team foundation server, user acceptance testing, ms sql server, asp net mvc, sql server reporting, sql stored procedures, visual studio 2010, software development methodologies,

Analyste programmeur	
Soft skills	Hard skills
developpement, applications, informatique, connaissance, solutions, analyse, web, environnement, ~~recherche~~, experience ...	net, sql, programmation, javascript, server, java, cgi, agile, oracle, langage ...
~~sql server~~, ~~bases donnees~~, esprit equipe, ~~mise place~~, resolution problemes, ~~base donnees~~, projets developpement, capacite analyse, ~~capacite travailler~~, developpement logiciel ...	asp net, vb net, html css, visual studio, ~~intelligence affaires~~, services web, applications web, apache tomcat, oracle sql, ms sql ...
~~ms visual studio~~, esprit analyse synthese,~~ms sql server~~, capacite travailler equipe, team foundation server, ~~sql server 2008~~, ~~asp net mvc~~, capacite travailler pression, ~~visual studio 2010~~ ...	master data management, object oriented programming, high pressure environment, visual basic net, visual studio team

Tableau 4. *Liste de compétences sous forme de n-grammes classées par unigrammes, bigrammes, trigrammes pour les métiers* programmer analyst *et* analyste programmeur *obtenue à partir de respectivement 185 et 169 offres. Les n-grammes considérés comme non pertinents ont été barrés. Les compétences sont triées en ordre décroissant de score.*

n-grammes	Anglais						Français					
	Unigramme		Bigramme		Trigramme		Unigramme		Bigramme		Trigramme	
	soft	hard	soft	hard	soft	hard	soft	hard	soft	hard	soft	hard
Total	60	60	60	60	57	60	60	60	60	60	42	9
Pertinents	48	50	53	58	57	59	57	52	52	56	33	8
Précision	0,80	0,83	0,88	0,97	1,0	0,98	0,95	0,87	0,87	0,93	0,79	0,88

Tableau 5. *Synthèse de l'évaluation des résultats.* soft *et* hard *représentent respectivement les compétences transversales et les compétences métier.*

Nous avons barré les n-grammes qui ont été considérés comme non pertinents par l'expert. Le tableau 5 présente une synthèse de l'évaluation de ces compétences en termes de *précision*. Ne disposant pas d'une liste complète des compétences pour chaque métier, nous n'avons pas été en mesure de calculer le *rappel*.

Même si l'échantillon évalué est de taille relativement petite (656 compétences sur environ 60 000 compétences générées), les résultats obtenus sont de très bonne qualité (0,89 sur l'ensemble de l'évaluation). La qualité des listes en anglais (0,91) est légèrement meilleure que celles en français (0,87). Nous attribuons cette différence au plus petit nombre d'offres d'emploi pour les métiers considérés en langue française (497 et 185 offres en anglais contre 127 et 169 offres en français). Ce nombre restreint d'offres d'emploi en français explique aussi le faible nombre de trigrammes obtenus pour le métier d'*analyste programmeur*.

L'analyse détaillée montre par ailleurs une présence importante de termes anglais présents dans les résultats français. Nous attribuons ce mélange à une utilisation fréquente de termes anglais dans les offres d'emploi en français. Celles-ci contiennent des ambiguïtés pour certaines compétences (telles que *leadership*, *management* ou encore *marketing*) ainsi que pour certaines fonctions (*manager*, *trader*, *designer*, etc.), ce qui complique la tâche de détection de langue. L'utilisation des n-grammes de mots occasionne par ailleurs des erreurs entre certaines compétences suivant la taille du n-gramme considéré (*access* et *access to a vehicule*, *office* et *office management*, etc.). Nous constatons aussi une redondance de certaines compétences en fonction de l'usage du singulier ou du pluriel (par exemple : *analyse financière* et *analyses financières*, *bases données* et *base données*) ou de versions différentes d'un logiciel (par exemple : *visual studio 2005/2010* ou *AutoCad 2015/2017*).

Le tableau 5 montre des résultats de meilleure qualité pour les bigrammes et les trigrammes que pour les unigrammes (des précisions moyennes de 0,86 pour les unigrammes, 0,91 pour les bigrammes et 0,93 pour les trigrammes). Nous travaillons à leur amélioration, car l'observation des offres d'emploi et des résultats montre que les technologies requises sont généralement présentées sous forme d'unigrammes (par exemple *cga*, *sap*, *Java*, *MySQL*, etc.). La mise en place du dictionnaire dynamique (section 5.5) a permis d'améliorer globalement la qualité des n-grammes obtenus, cependant des termes ou expressions courants dans les offres d'emploi et dans le dictionnaire viennent parfois bruiter les résultats (par exemple *assurance*). À partir des données du tableau 5, on peut calculer que la précision moyenne des compétences tech-

niques (*hard skills*) 0,92 est légèrement meilleure que celle des compétences transversales (*soft skills*) qui est de 0,88. Afin d'analyser cette différence, nous avons effectué une évaluation de la qualité du dictionnaire généré pour les compétences transversales et les résultats montrent qu'une partie du vocabulaire contenu dans le dictionnaire est non pertinente (précision de 0,87 sur 250 compétences), particulièrement les unigrammes (0,60), ce qui se répercute sur la qualité des résultats finaux.

7. Conclusion et travaux futurs

Nous avons présenté dans cet article les travaux réalisés sur la génération automatique de ressources linguistiques pour les besoins de l'e-recrutement. À partir de dix millions de profils issus de plusieurs réseaux sociaux, ainsi que 45 000 offres d'emploi sur les 300 000 récoltées sur Internet, nous avons fait émerger des compétences et des connaissances communes pour construire un ensemble structuré des termes et concepts représentant chaque métier avec ses compétences associées. Ces offres contenant le profil minimal recherché pour un métier ont été utilisées afin de détecter les compétences. Ces offres d'emploi étant collectées sur Internet, le modèle pourra ainsi s'enrichir de façon semi-automatique avec l'apparition et la disparition de métiers. Nous avons présenté par la suite les premiers résultats obtenus ainsi qu'une évaluation manuelle réalisée par un expert du domaine. L'analyse détaillée a montré des résultats de très bonne qualité (précision moyenne de 0,89) particulièrement en anglais. Nous attribuons cette différence principalement à la quantité plus faible d'offres d'emploi pour les métiers considérés en langue française.

Nous envisageons des traitements plus fins afin de regrouper les compétences fortement similaires ou qui ne diffèrent que par de faibles variations d'écriture. Nous souhaitons utiliser cette ontologie afin d'évaluer si un candidat dispose de toutes les compétences pour un métier ou encore au travers d'un outil de génération de texte afin de suggérer à un candidat comment mettre en avant ses compétences ou encore suggérer des mots-clés pour les recruteurs qui effectuent des recherches dans des bases de profils de candidats. Nous prévoyons par ailleurs de continuer à augmenter la taille de l'ensemble des offres d'emploi afin de couvrir un nombre de métiers plus important pour améliorer la qualité des résultats. Le processus complet étant automatisé, il est possible, et nous l'avons d'ailleurs expérimenté à quelques reprises, de créer une nouvelle ontologie aussitôt que de nouvelles offres d'emploi deviennent disponibles. Il serait toutefois intéressant d'effectuer une mise à jour de l'ontologie existante.

Par ailleurs, nous envisageons l'utilisation de notre système en complément d'un processus classique de sélection des candidats *actifs*, c'est-à-dire ayant posé leur candidature, nous pourrions ainsi identifier les candidats *passifs*, qui ne sont pas en recherche d'emploi, mais qui pourraient être intéressés par de nouvelles opportunités, en parcourant les différents médias sociaux afin de récolter les profils les plus en adéquation avec une offre d'emploi. Une fois cette collecte terminée, une proposition d'opportunité serait transmise à ces candidats *passifs* qui pourraient alors décider si la proposition les intéresse suffisamment pour postuler.

Remerciements

Les auteurs tiennent à remercier le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), qui a financé ces travaux dans le cadre d'une subvention de recherche et développement coopérative ainsi que les membres de l'équipe du projet BPP, en particulier Fabrizio Gotti pour ses contributions sans oublier Éric Tondo et Ludovic Bourg pour leur connaissance du domaine et les discussions fructueuses. L'article a été finalisé dans le cadre d'un séjour de Guy Lapalme à l'IMERA (université Aix-Marseille).

8. Bibliographie

Bendaoud R., Rouane Hacene A. M., Toussaint Y., Delecroix B., Napoli A., « Construction d'une ontologie à partir d'un corpus de textes avec l'ACF », *IC 2007*, Grenoble, France, 2007.

Colucci S., Di Noia T., Di Sciascio E., Donini F. M., Mongiello M., Mottola M., « A formal approach to ontology-based semantic match of skills descriptions », *J. UCS*, vol. 9, n° 12, p. 1437-1454, 2003.

Colucci S., Di Noia T., Di Sciascio E., Donini F. M., Ragone A., Trizio M., « A Semantic-based Search Engine for Professional Knowledge », *Proc. 7th Int. Conf. on Knowledge Management and Knowledge Technologies (I-KNOW 2007),(Sep 2007)*, p. 472-475, 2007.

Colucci S., Tinelli E., Giannini S., Di Sciascio E., Donini F. M., « Knowledge Compilation for Core Competence Extraction in Organizations », *Business Information Systems*, Springer, p. 163-174, 2013.

Desmontils E., Jacquin C., Morin E., « Indexation sémantique de documents sur le Web : application aux ressources humaines », *Journées de l'AS-CNRS Web Sémantique*, 2002.

Gómez-Pérez A., Ramírez J., Villazón-Terrazas B., « An ontology for modelling human resources management based on standards », *Knowledge-Based Intelligent Information and Engineering Systems*, Springer, p. 534-541, 2007.

Hoffart J., Suchanek F. M., Berberich K., Lewis-Kelham E., de Melo G., Weikum G., « YAGO2 : Exploring and Querying World Knowledge in Time, Space, Context, and Many Languages », *Proceedings of the 20th International Conference Companion on World Wide Web*, p. 229-232, 2011.

Kessler R., Béchet N., Roche M., El-Bèze M., Torres-Moreno J. M., « Automatic profiling system for ranking candidates answers in human resources », *On the Move to Meaningful Internet Systems : OTM 2008 Workshops*, Springer, p. 625-634, 2008.

Kessler R., Lapalme G., Tondo É., « Génération d'une ontologie dans le domaine des ressources humaines », *CORIA 2016*, Toulouse, 03/2016, 2016.

Kmail A. B., Maree M., Belkhatir M., Alhashmi S. M., « An Automatic Online Recruitment System Based on Exploiting Multiple Semantic Resources and Concept-Relatedness Measures », *Tools with Artificial Intelligence (ICTAI), 2015 IEEE 27th International Conference on*, p. 620-627, 2015.

Lau T., Sure Y., « Introducing ontology-based skills management at a large insurance company », *Proceedings of the Modellierung 2002*p. 123-134, 2002.

le Vrang M., Papantoniou A., Pauwels E., Fannes P., Vandensteen D., De Smedt J., « ESCO : Boosting Job Matching in Europe with Semantic Interoperability », *Computer*, vol. 47, nº 10, p. 57-64, Oct, 2014.

Levenshtein V. I., « Binary codes capable of correcting deletions, insertions and reversals. », *Soviet Physics Doklady*, vol. 10, nº 8, p. 707-710, 1966.

Loth R., Battistelli D., Chaumartin F.-R., De Mazancourt H., Minel J.-L., Vinckx A., « Linguistic information extraction for job ads (SIRE project) », *Adaptivity, Personalization and Fusion of Heterogeneous Information*, p. 222-224, 2010.

Miles A., Bechhofer S., « SKOS Simple Knowledge Organization System Reference », 2009.

Mochol M., Simperl E. P. B., « Practical guidelines for building semantic erecruitment applications », *International Conference on Knowledge Management, Special Track : Advanced Semantic Technologies (AST'06)*, Citeseer, 2006.

Osborne J. D., Flatow J., Holko M., Lin S. M., Kibbe W. A., Zhu L. J., Danila M. I., Feng G., Chisholm R. L., « Annotating the human genome with Disease Ontology », *BMC Genomics*, 2009.

Roche M., Kodratoff Y., « Pruning terminology extracted from a specialized corpus for CV ontology acquisition », *On the Move to Meaningful Internet Systems 2006 : OTM 2006 Workshops*, Springer, p. 1107-1116, 2006.

Sami G., Béchet N., Berio G., « Ontologies from Textual Resources : A Pattern Based Improvement Using Deep Linguistic Information », *Workshop on on Ontology and Semantic Web Patterns (WOP)*, vol. 1302 of *Proceedings of Workshop on on Ontology and Semantic Web Patterns (WOP)*, p. 14-25, 2014.

Sivabalan L., Yazdanifard R., Ismail N. H., « How to Transform the Traditional Way of Recruitment into Online System », *International Business Research*, vol. 7, p. 178, 2014.

Trichet F., Bourse M., Leclerc M., Morin E., *Human Resource Management and Semantic Web Technologies*, springer edn, ICTTA, Berlin, 2004.

Trog D., Christiaens S., Zhao G., de Laaf J., « Toward a Community Vision Driven Topical Ontology in Human Resource Management », *On the Move to Meaningful Internet Systems : OTM 2008 Workshops*, Springer, p. 615-624, 2008.

Tétreault M., Dufresne A., Gagnon M., « Development of an Ontology-Based E-Recruitment Application that Integrates Social Web », *Electronic business interoperability : Concepts, opportunities and challenges*p. 363-395, 2011.

Yahiaoui L., Boufaïda Z., Prié Y., « Automatisation du e-recrutement dans le cadre du web sémantique », *Journée francophones d'Ingénierie des Connaissances, IC'2006*, 2006.

Zimmermann T., Kotschenreuther L., Schmidt K., « Data-driven HR - Résumé Analysis Based on Natural Language Processing and Machine Learning », *CoRR*, 2016.