Génération d'une ontologie dans le domaine des ressources humaines

Rémy Kessler* — **Guy Lapalme*** — **Eric Tondo****

* *RALI - Département d'informatique et de recherche opérationnelle*
Université de Montréal
C.P. 6128, Succ Centre-Ville, Montréal, Québec, Canada H3C 3J7
{kessler,lapalme}@iro.umontreal.ca
** *LittleBigJob*
204 rue du St-Sacrement, Montréal, Québec, Canada H2Y 1W8
eric.tondo@littlebigjob.com

RÉSUMÉ. *Notre époque est de plus en plus influencée par la prééminence des données intelligentes et du web sémantique. Les processus de recrutement n'en sont pas toujours facilités en particulier en matière de recherche de profils et de talents. La plupart des systèmes d'appariement entre une offre d'emploi et un profil s'appuient sur une ou plusieurs ressources linguistiques, mais se heurtent à la difficulté de développer et à entretenir des ressources spécifiques à chaque domaine. Nous proposons une approche de construction semi-automatique d'une ontologie s'appuyant sur des millions de profils issus de plusieurs réseaux sociaux et sur des dizaines de milliers d'offres d'emploi collectées sur internet. Une évaluation manuelle a été réalisée par un expert du domaine et a montré des résultats de très bonne qualité.*

ABSTRACT. *The current era is increasingly influenced by the prominence of smart data and the semantic web. This does not make recruitment processes easier, especially when seeking candidates and skills. Most systems matching job offers to profiles build on one or more language resources. Developing and maintaining specific resources in each field is difficult. From millions of user profiles gathered on multiple social networks and tens of thousands of job offers collected on the internet, we describe an approach to generating a semi-automatic ontology. A manual evaluation of the results has shown excellent results.*

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

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

1. Introduction

Durant la dernière décennie, nous avons assisté à un développement rapide du web et des réseaux sociaux qui a modifié la dynamique de la recherche d'emploi. 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 éléments recueillis sur un individu à un premier niveau de recherche sont peu structurés, disparates, incomplets, redondants, parfois obsolètes et peuvent être biaisés, voire trompeurs (ex. : à cause des homonymes).

Dans le cadre du projet Butterfly Predictive Project[2], nous développons une plateforme pour améliorer de manière importante la mise en correspondance entre des candidats et des offres d'emploi en étudiant comment les développements récents et à venir du web peuvent aider résoudre ces problématiques. Nous faisons l'hypothèse que l'acquisition et l'exploitation de cette trace laissée par les individus, diffuse, plus ou moins accessible et peu structurée sont une voie prometteuse pour l'expression du positionnement professionnel des candidats, pour la caractérisation des emplois et pour leur mise en correspondance. En complément du processus habituel de sélection d'un sous-ensemble des candidats actifs, c'est-à-dire ayant posé leur candidature, nous proposons d'identifier des 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 être intégrés à la liste de candidats potentiels pour le poste. Même si de grands efforts ont été déployés ces dernières années afin de développer des ressources linguistiques afin d'aider les systèmes d'appariement de candidatures et d'offres d'emploi, ces derniers se heurtent à la difficulté de créer des ontologies/taxonomies spécifiques à chaque domaine et particulièrement à leur entretien et 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 entraine l'apparition de nouveaux métiers et de nouvelles compétences ainsi que la disparition d'autres. Même si chaque profil est différent dans le détail, il

1. http://entreprise.regionsjob.com/enquetes/reseaux_sociaux/resultats_enquete_2.pdf
2. http://rali.iro.umontreal.ca/rali/?q=fr/butterfly-predictive-project

semble cependant envisageable qu'un regroupement d'informations issues des mêmes métiers fasse émerger des relations communes permettant de construire un ensemble structuré des termes et concepts représentant chaque domaine. 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. évaluer si un candidat a toutes les compétences pour un métier), de la génération de texte (p. ex. suggérer à un candidat comment mettre en avant ses compétences les plus en adéquation pour un poste).

2. Travaux connexes

De nos jours, des 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 est une tâche difficile à accomplir. De grands efforts ont cependant été déployés ces dernières années afin de constituer des ressources linguistiques afin d'améliorer les systèmes d'appariement.

La première étude a été proposée par (Lau et Sure, 2002) : ils décrivent une méthodologie afin de développer une ontologie (en se basant sur une étude de cas de la société Swiss Life) centrée sur le domaine des technologies de l'information (TI). Ils précisent que celle-ci a 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 candidature et d'offres d'emploi à l'aide d'une ontologie ont été proposés par (Colucci *et al.*, 2003 ; Colucci *et al.*, 2007). Leur système, IMPAKT[3](Colucci *et al.*, 2013) combine une approche sémantique avec une ontologie et repose sur des méthodes de formalisation des raisonnements et des connaissances. Le système propose un appariement en effectuant des correspondances partielles ou complètes des compétences entre les offres d'emplois et les candidatures. La spécificité des informations contenues dans les documents d'une candidature et d'une offre d'emploi a conduit au développement d'approches sémantiques pour construire des ressources linguistiques.

(Desmontils *et al.*, 2002 ; Trichet *et al.*, 2004) proposent une méthode d'indexation sémantique. La méthode consiste à exploiter les caractéristiques dispositionnelles du document afin d'identifier chacune des parties et l'indexer en conséquence. Ils concluent les travaux en proposant une instanciation d'ontologie en partant des données récoltées et en s'appuyant sur des ressources externes (base ROME[4] et CIGREF[5]). Même si l'approche semble intéressante, l'absence de résultat ne permet pas d'évaluer l'apport de cette indexation particulière. (Mochol et Simperl, 2006) décrivent l'importance d'une ontologie commune (HR ontology) ainsi qu'un guide pour

3. Information Management and Processing with the Aid of Knowledge-based Technologies
4. Répertoire Organisationnel des Métiers et Emplois
5. Club Informatique des Grandes Entreprises Françaises

mettre en place ce type d'application tandis que (Trichet *et al.*, 2004) décrivent différentes approches permettant la gestion des compétences à l'aide d'ontologies dans le cadre du e-recrutement.

Dans le cadre du projet Prolix, (Trog *et al.*, 2008) décrivent une ontologie de ressources humaines basée sur le cas de British Telecom. Ils proposent une architecture en plusieurs niveaux en fonction des compétences, des interactions et du contexte. Toujours dans l'idée de créer des ressources linguistiques (Gómez-Pérez *et al.*, 2007) proposent une annotation sémantique des documents (offre d'emploi et CV) afin de construire différentes ontologies. Développées en anglais et disponibles en ligne[6], ces ontologies décrivent des compétences et formations spécifiques au domaine des TI tandis que d'autres sont plus généralistes comme les classes de métiers, permis de conduire ou encore les secteurs d'activités.

(Roche et Kodratoff, 2006) décrivent une étude d'extraction de terminologie spécifique sur un corpus de CV. Leur approche permet d'extraire 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 de les classer 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 la maquette d'une plateforme de recrutement intégrant les technologies du web sémantique ainsi que l'élaboration d'une ontologie à l'aide d'OWL dédiée au domaine TI. L'approche décrit les avantages d'une application de ce type ainsi que différents scénarios de recrutement.

Plus récemment, (le Vrang *et al.*, 2014) présente l'ontologie ESCO[7], un projet européen multilingue permettant de classifier compétences, métiers et 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 limite les domaines et métiers couverts. 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.

Ces travaux montrent que différentes approches ont été envisagées, que ce soit avec des méthodes statistiques, sémantiques ou encore à l'aide de ressources linguistiques complémentaires. Les réseaux sociaux et plus généralement l'information récoltée sur internet sur les candidats peuvent être une source pertinente afin de faire émerger des compétences et des connaissances communes pour construire un ensemble structuré de termes et de concepts.

---

6. `http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/99-hrmontology`
7. European Skills Competences and Occupations `https://ec.europa.eu/esco/`

3. Données et connaissances

Le projet BPP s'appuie sur une modélisation en termes de 44 secteurs d'activités (univers), chacun regroupant des familles de métiers assez proches et un ou plusieurs univers connexes. Les univers connexes au secteur considéré sont ceux pour lesquels une transition est envisageable pour un profil envisageant un changement de secteur. Par exemple, un profil issu de l'univers *Conseil et service informatiques, édition de logiciels* pourra plus facilement envisager une transition dans les domaines *Télécoms, Hébergement, Internet* ou *Ingénierie - R&D* que dans celui de *la cosmétique*. Nous avons combiné ces connaissances avec l'ontologie issue du modèle ESCO, ainsi qu'un dictionnaire des métiers issus de la Classification nationale des professions[8] (CNP) et du Répertoire opérationnel des métiers et des emplois (ROME)[9]. 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'emplois (anglais/français) en 500 profils de groupes professionnels tandis que le ROME comprend 10 000 appellations différentes réparties en 531 groupes.

Par ailleurs, plus de 10 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 regroupe différentes informations sur le parcours du candidat tels que ses diplômes et formations ainsi que la date d'obtention de chacun d'eux. De la même façon, une section du profil contient ses expériences. Chacune contenant 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 de 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, ses langues maitrisées, ses loisirs, le nombre de relations ou encore les compétences acquises au cours de son parcours professionnel. 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 compétences issues de ces profils, environ 2,3 millions.

En complément de ces données, 100 000 offres d'emplois ont été collectées sur internet, 60 000 en anglais et 40 000 en français. Ces offres d'emplois couvrent un grand nombre de métiers et sont issues, elles aussi, du Canada ou de la France. Chaque offre

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

	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*

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 d'informations intéressantes afin de constituer une ontologie de façon semi-automatique.

4. Méthodologie

La plupart des systèmes d'appariement entre une offre d'emploi et un profil présentés dans la section 2 s'appuient sur une ou plusieurs ressources linguistiques, de type ontologie ou taxonomie, mais se heurtent cependant à la difficulté de développer des ressources spécifiques à chaque domaine, ainsi qu'à leur entretien. Afin de pallier ce problème, nous souhaitons développer un système permettant de générer une ontologie de façon semi-automatique en s'appuyant sur les données collectées sur internet et présentées dans la section précédente. Nous faisons l'hypothèse que l'exploitation de ces informations est une voie prometteuse pour constituer un référentiel commun, une représentation de chaque domaine avec des liens logiques.

Dans un premier temps, nous avons effectué une comparaison des compétences issues des réseaux sociaux et les compétences contenues dans l'ontologie ESCO. Les résultats ont montré de nombreuses correspondances exactes ou partielles (plus d'un million) et variées. Par ailleurs 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 modeling*, *food cost management* pour les compétences).

Par ailleurs, il est impossible 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 sénior en technologies de l'information cumulera les 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 limiter à cette source comme identification des compétences requises pour un emploi.

Afin de pallier ces problèmes, nous avons décidé d'utiliser comme source de documents la collection d'offres d'emplois. En effet, celles-ci définissent les expériences et qualifications recherchées pour un métier particulier. Comme nous disposons d'un nombre important d'offres d'emplois (100 000), nous croyons être en mesure de faire émerger les compétences et connaissances communes pour construire un ensemble structuré des termes et concepts représentant chaque métier. L'ensemble des offres d'emplois étant collecté sur internet, le modèle s'enrichira à terme de façon semi-automatique avec l'apparition et disparition de métiers dans les offres d'emplois.

5. Vue d'ensemble du système AGOHRA

La figure 1 présente une vue d'ensemble du système AGOHRA[10]. Au cours d'une première étape ①, on effectue une normalisation des offres d'emploi. Le module suivant ② utilise les titres des offres d'emplois afin de détecter les métiers avant d'y associer l'univers concerné. L'étape suivante ③ consiste à utiliser l'ensemble du vocabulaire récolté afin de faire émerger les compétences. Le dernier module ④ transforme ensuite les informations ordonnées et structurées en une ontologie au format RDF.

5.1. *Normalisation des offres d'emploi*

Au cours d'une première étape, nous effectuons une séparation des 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 la collection 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'emplois afin d'éviter d'augmenter artificiellement les fréquences des termes. Environ 6000 offres d'emplois sont ainsi écartées de la collection. En parallèle avec cette étape, une normalisation des titres des offres d'emplois est effectuée afin de détecter les métiers en utilisant différents patrons afin d'améliorer la qualité du processus. On effectue ainsi une suppression des accents, des caractères particuliers tels que « - », « / », « ( ) »ou encore des expressions couramment utilisées dans les titres d'offres d'emplois tels 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

10. Automatic Generation of an Ontology for Human Resource Applications

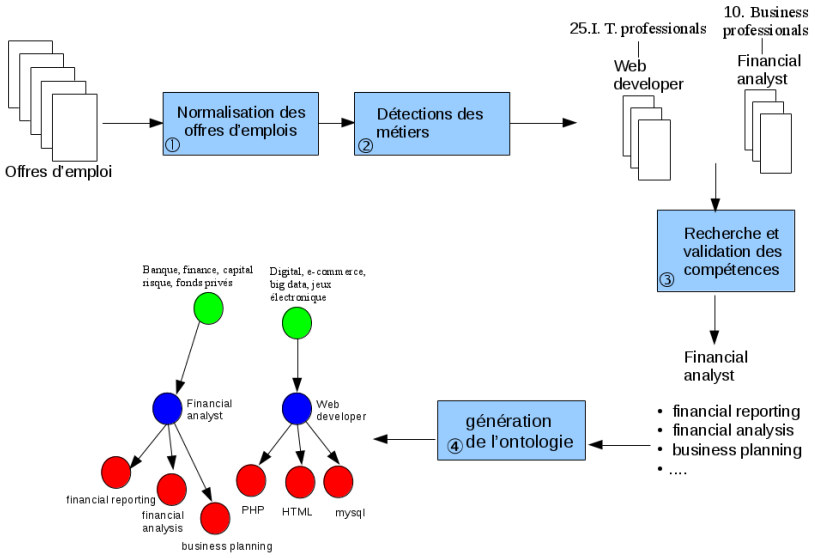

Figure 1. *Vue d'ensemble du système AGOHRA.*

nombres (numériques et/ou textuelles), les symboles spéciaux ainsi que les termes contenus dans un anti-dictionnaire spécifique et adapté à notre problème.

5.2. *Détection des métiers*

Nous utilisons au cours de cette étape les titres des offres d'emplois normalisées (voir 5.1) afin de construire une liste de métiers. Diverses règles sont ainsi appliquées afin d'extraire les métiers contenus dans ces titres. À l'aide d'un seuil de fréquence minimum déterminé empiriquement, nous ne conservons que les métiers avec une fréquence importante. Une première version du système permettait de générer les résultats pour l'ensemble des métiers, mais nous avons décidé de privilégier les métiers ciblés par notre partenaire industriel afin de filtrer les métiers avec peu de qualifications requises, comme *gardienne de chat*, *nounous*, etc. Nous filtrons ainsi une grande quantité d'offres d'emploi pour n'en conserver qu'environ 9 000 (6 000 en anglais et 3 000 en français).

5.3. *Recherche et validation des compétences*

L'objectif de cette étape (③) est d'utiliser le vocabulaire issu des offres d'emplois 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 souligne (Loth *et al.*, 2010). (Kessler *et al.*, 2008) montre 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, mais que celles-ci suivent cependant un ordre conventionnel dans leur apparition. 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 la collection de profils de réseaux sociaux décrite en section 3 montre que plus de 80% des compétences se présentent sous la forme de n-grammes de mots (par exemple *financial modeling*, *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% de vide. Compte tenu de ces observations, nous avons décidé de transformer l'ensemble du vocabulaire issu des offres d'emplois sous forme d'unigrammes, de 2-grammes et 3-grammes et de les ordonner selon un score S_{job}, inspiré du *TF-IDF* :

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

avec u l'unité lexicale considérée (unigramme, 2-grammes ou 3-grammes), $tf(u)$ 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 certains termes et expressions courants dans les offres d'emplois et qui viennent parfois bruiter les résultats (par exemple *employment equity*, *experience working*), nous effectuons par la suite une comparaison de la liste ordonnée de n-grammes obtenus avec un arbre préfixe contenant environ 25 000 compétences issues des profils de réseaux sociaux. Cette structure permet d'effectuer des comparaisons rapides entre l'arbre de compétences et les n-grammes candidats tout en conservant les variations d'écriture pour une même compétence (p. ex. *en charge de projet*, *en charge de projets*, *conduite de projets*, etc.).

5.4. *Génération de l'ontologie*

Une fois l'ensemble des comparaisons effectuées, nous générons une structure de données hiérarchique associant à chaque univers (`uni:n` dans la figure 2) par la relation `bpp:comprises` un ensemble de métiers `occ:i` nommés *occupations* par

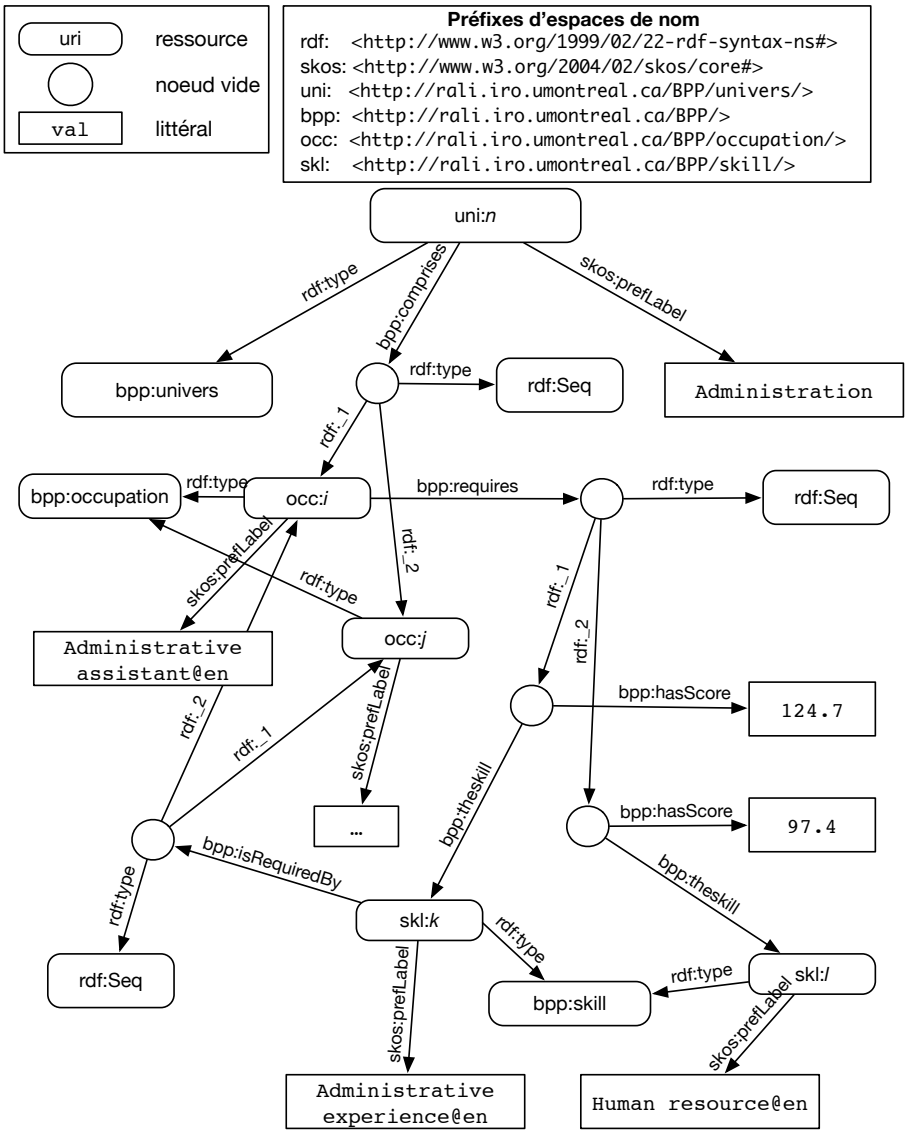

Figure 2. *Organisation du graphe RDF de l'ontologie générée.*

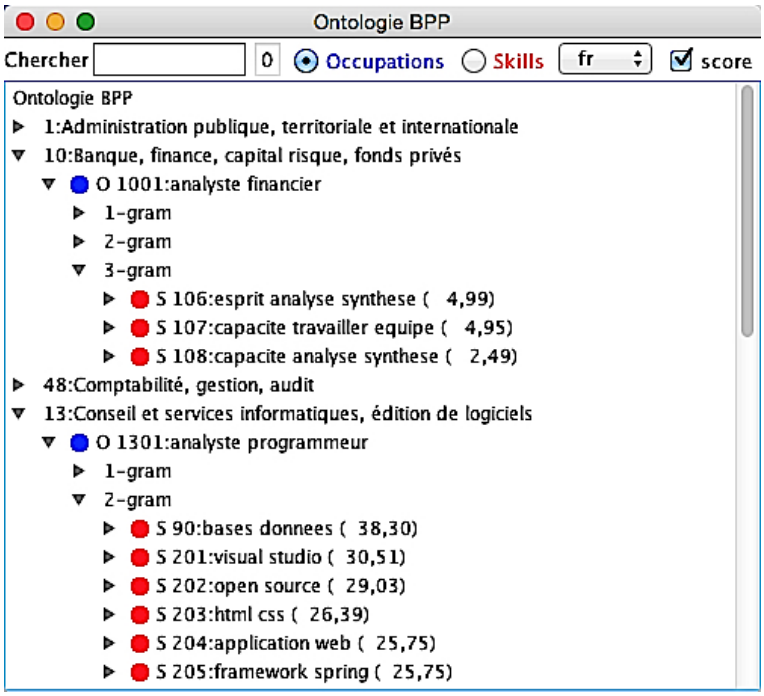

Figure 3. *Browser affichant un extrait de l'ontologie. Chaque concept présent dans l'ontologie est déterminé à l'aide d'un identifiant numérique, les compétences sont précédées de la lettre S (Skills) les métiers par O (Occupations), suivi du numéro associé à l'univers (par exemple O1301 analyste programmeur appartient à l'univers 13* Conseil et services informatiques, édition de logiciels*).*

souci de compatibilité avec la nomenclature ESCO. Chaque occupation est liée par la relation `bpp:requires` à des listes de compétences (*skills* dans ESCO). 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. Le réseau RDF pour l'anglais contient 25 569 triplets pour 57 métiers et 1608 compétences et celui pour le français comprend 4 183 triplets pour 11 métiers et 330 compétences. Ils sont interrogeables avec SPARQL dans nos applications. Pour en faciliter l'exploration visuelle, nous avons aussi développé un browser spécialisé qui affiche la structure RDF en faisant ressortir les différents niveaux : univers, métiers et compétences. La figure 3 présente une capture d'écran contenant un extrait de l'ontologie finale avec les métiers évalués en section 6.

financial analyst	
unigramme	accounting, finance, financial, cma, budget, analysis, ~~end~~, review, reporting, projects, support, ~~including~~, process, ~~key~~, business, ~~com~~, ~~canada~~, information, ~~knowledge~~, ~~maintain~~, develop, ~~qualifications~~, responsibilities, ~~requirements~~, ~~environment~~, team, work, ~~strong~~, ~~working~~, ~~experience~~
bigrammes	financial analyst, financial analysis, financial reporting, financial modeling, journal entries, professional accounting, variance analysis, accounting finance, financial planning, finance accounting, balance sheet, management reporting, financial reports, financial statements, finance business, ~~special projects~~, financial systems, management reports, ~~internal external~~, analytical skills, ~~ability work~~, financial services, strong analytical, ~~experience working~~, strong organizational, business processes, decision making, communication skills, ~~skills strong~~, fast paced
trigrammes	financial planning analysis, analytical problem solving, problem solving skills, advanced excel skills, communication interpersonal skills, verbal written communication, ability work independently, organizational time management, strong analytical skills, fast paced environment, ability multi task, oral written communication, exceptional interpersonal skills, financial analysis reporting, verbal communication skills, time management skills, strong communication skills, excel pivot tables, written oral communication, strong interpersonal skills, written verbal communication, ability work pressure, knowledge microsoft office, microsoft office suite, key performance indicators, work team environment, strong organizational skills, excellent interpersonal skills, excellent communication skills, interpersonal communication skills

analyste financier	
unigramme	finances, financement, analyse, ~~social~~, finance, ~~public~~, ~~documents~~, ~~direction~~, process, rigueur, audit, ~~experience~~, office, procedures, ~~realisation~~, business, ~~connaissance~~, support, ~~formation~~, organisation, gestion, service, ~~premiere~~, ~~international~~, commerce, reporting, responsable, developpement, ~~services~~, ~~clients~~
bigrammes	vision strategique, tableaux bord, strategique gestion, strategie commerciale, sens organisation, sens analyse, ~~ressources humaines~~, processus budgetaire, politiques publiques, optimisation processus, modelisation financiere, ~~mise place~~, middle office, gestion risque, gestion projets, gestion priorites, gestion groupe, excel vba, etats financiers, ~~direction generale~~, developpement local, developpement economique, creation entreprise, controle gestion, business plan, bases donnees, analyses financieres, analyse risques, analyse financiere, analyse economique
trigrammes	esprit analyse synthese, capacite travailler equipe, capacite analyse synthese

Tableau 2. *Liste de compétences sous forme de n-grammes pour les métiers "financial analyst" et "analyste financier" obtenues à partir de respectivement 135 et 44 offres.*

programmer analyst	
unigramme	programmer, analyst, programming, net, application, test, applications, developing, support, technical, systems, design, development, tools, develop, data, ~~knowledge~~, ~~following~~, solutions, environment, maintain, requirements, professional, ~~including~~, ~~work~~, business, qualifications, team, ~~experience~~, ~~time~~
bigrammes	sql server, application development, web applications, programming languages, systems development, information technology, web services, unit testing, working knowledge, business requirements, working experience, crystal reports, visual studio, software development, ~~latin america~~, design development, technology solutions, version control, development team, database design, production support, web development, operating systems, best practices, wealth management, business systems, ~~general knowledge~~, ~~experience working~~, ~~internal external~~, communication skills
trigrammes	object oriented programming, microsoft sql server, problem solving skills, sql reporting services, java enterprise edition, corporate social responsibility, team foundation server, sql server reporting, ability work independently, microsoft dynamics crm, ~~job description development~~, ~~visual source safe~~, enterprise content management, excellent communication skills, service oriented architecture, enterprise application integration, excellent organizational skills, customer service skills, able work independently, written verbal communication, analytical problem solving, software development lifecycle, build strong relationships, user acceptance testing, problem solving ability, strong analytical skills, work minimal supervision, written oral communication, strong interpersonal skills, communication interpersonal skills

analyste programmeur	
unigramme	applications, java, ~~environnement~~, net, informatique, javascript, ingenierie, developpement, programmation, ~~end~~, conception, mysql, ~~connaissance~~, windows, maintenance, solutions, ~~experience~~, ~~competences~~, ~~information~~, ~~techniques~~, ~~direction~~, ~~formation~~, ~~accompagnement~~, ~~service~~, analyse, anglais, autonome, responsable, ~~services~~, ~~clients~~
bigrammes	bases donnees, visual studio, open source, html css, application web, framework spring, apache tomcat, esprit equipe, ~~mise place~~, ~~enseignement superieur~~, systeme information, base donnees, sql server, ~~direction generale~~, rpg iii, analyse besoins, suivi production, zend studio, administration reseau, ~~propriete intellectuelle~~, ~~parle ecrit~~, developpement logiciel, nouvelles technologies, maitrise anglais, esprit synthese, dynamics crm, sens analyse, capacite analyse, ~~supply chain~~, tableaux bord,
trigrammes	microsoft sql server, esprit analyse synthese, capacite travailler equipe,

Tableau 3. *Liste de compétences sous forme de n-grammes pour les métiers "programmer analyst" et "analyste programmeur" obtenues à partir de respectivement 112 et 69 offres.*

	anglais			français		
	uni.	bi.	tri.	uni.	bi.	tri.
N-grammes total	60	60	60	60	60	6
N-grammes pertinents	42	51	58	34	51	6
Précision	0.7	0.85	0.96	0.56	0.85	1

Tableau 4. *synthèse de l'évaluation des résultats.*

6. Résultats

Nous présentons ici les compétences en ordre décroissant de S_{job} (voir section 5.3) pour les métiers *analyste financier* (Tableau 2) et *analyste programmeur* (Tableau 3) en français et en anglais. Cette portion de l'ontologie a été évaluée par un expert du domaine. Nous avons ~~barré~~ les n-grammes qui ont été considérés comme non pertinents par celui-ci. Le tableau 4 présente une synthèse de cette évaluation en termes de *précision*. Ne disposant pas de liste de l'ensemble des compétences pour un métier, nous n'avons pas été en mesure de calculer le *rappel*.

Même si l'échantillon évalué est de taille relativement petite, les résultats obtenus sont de très bonne qualité (0.79 sur l'ensemble de l'évaluation). La qualité des listes en anglais (0.83) est meilleure que celles en français (0.75). Nous attribuons cette différence au plus petit nombre d'offres d'emploi pour les métiers considérés en langue française (135 et 112 offres en anglais contre 44 et 69 offres en français). Ce nombre restreint d'offres d'emploi en français explique aussi le faible nombre de trigrammes obtenus pour les métiers d'*analyste financier* et d'*analyste programmeur*.

Le tableau 4 montre des résultats de meilleure qualité pour les bigrammes et trigrammes que pour les unigrammes (respectivement 0.63 pour les unigrammes, 0.85 pour les bigrammes et 0.97 pour les trigrammes), quelle que soit la langue considérée.

Nous observons cependant un certain nombre de compétences transversales (tels que *verbal/written communication skills*, *capacité à travailler en équipe*, etc.) qui peuvent être considérées pertinentes quelque soit le métier considéré. Nous envisageons des traitements particuliers afin d'être capable de les distinguer des compétences plus techniques, communément appelées *hard skills*[11]. Nous constatons par ailleurs une redondance de certaines compétences en fonction de l'usage du singulier ou du pluriel (par exemple : *analyse financière* et *analyses financières*, *bases donnees et base donnees*). De la même façon, on observe une fréquence importante de certaines compétences (par exemple *excel*, *microsoft office*, etc.) qui se retrouvent dans un grand nombre de métiers ainsi que certain termes/expressions courants dans les offres d'emplois et qui viennent parfois bruiter les résultats (par exemple *employment equity*, *human ressources*, *strong experience required*, etc.).

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

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 10 millions de profils issus de plusieurs réseaux sociaux, ainsi que 100 000 offres d'emplois 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. Ces offres contenant le profil minimal recherché pour un métier ont été utilisées afin de détecter les compétences associées. 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 bonnes qualités (précision moyenne de 0,79) particulièrement en anglais. Nous attribuons cette différence à la quantité plus faible d'offres d'emploi pour les métiers considérés en langue française.

Nous envisageons des traitements plus fins afin de distinguer les compétences transversales des compétences spécifiques à chaque métier. Nous souhaitons utiliser cette ontologie afin d'évaluer si un candidat dispose 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 la collection d'offres d'emplois afin de couvrir un nombre de métiers plus important pour améliorer la qualité des résultats.

Remerciements

Les auteurs tiennent à remercier le Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG), qui a partiellement financé ces travaux dans le cadre d'une subvention de recherche et développement coopérative.

8. Bibliographie

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.

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.

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, n^{o} 10, p. 57-64, Oct, 2014.

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.

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.

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.

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.