European Training Foundation

BIG DATA POUR LES SYSTÈMES D'INFORMATION/DE RENSEIGNEMENT SUR LE MARCHÉ DU TRAVAIL

BREF GUIDE D'INTRODUCTION

2019

Mario Mezzanzanica et *Fabio Mercorio*
Université de Milan-Bicocca – Centre de recherche CRISP, Italie

European Training Foundation

PREFACE

Ce guide d'introduction succinct aborde les principaux aspects conceptuels, méthodologiques et organisationnels de l'utilisation des Big Data pour l'information/le renseignement sur le marché du travail (IMT). Les lecteurs/utilisateurs cibles sont les statisticiens, les chercheurs, les analystes politiques et les décideurs des pays partenaires de la Fondation européenne pour la formation (ETF) qui sont confrontés aux défis de l'anticipation et de la diffusion d'indications sur la dynamique de la demande d'emplois, les compétences et les qualifications.

Les Big Data sont omniprésentes, mais leur potentiel et les façons dont elles peuvent être utilisées dans la recherche sociétale restent une nouveauté pour de nombreuses institutions publiques et parties prenantes dans les pays partenaires de l'ETF et les autres pays.

Ce guide d'introduction clarifie la manière dont les Big Data peuvent être utilisées pour aller au-delà des approches conventionnelles des systèmes d'information sur le marché du travail (SIMT) et apporter une valeur ajoutée aux statistiques établies. Les approches traditionnelles de diffusion de l'IMT, essentiellement basées sur des enquêtes, contiennent d'importants écueils: le coût, la rapidité, l'exactitude, l'usage, l'intégration et la couverture. Ces difficultés sont surmontables, mais leur résolution nécessitera d'appeler l'attention des gouvernements, des parties prenantes et de leurs partenaires donateurs.

Les sources et les analyses des Big Data contribuent à compléter et à enrichir les statistiques établies. L'analyse des Big Data peut être utilisée pour recenser les compétences par métiers, identifier les disparités de compétences, identifier les compétences obsolètes, réaliser une analyse prédictive de la demande de nouveaux métiers et de nouvelles compétences – en temps quasi réel. L'analyse des Big Data permet d'obtenir en temps réel des indications plus détaillées (granulaires), y compris une localisation géographique, ainsi qu'une analyse prédictive.

Le volume, la variété et la vélocité des Big Data vont continuer à augmenter. De grandes quantités de données numériques sont générées par les personnes, les organisations, les capteurs intelligents, les satellites, les caméras de surveillance, internet et d'innombrables autres dispositifs. L'idée même de donner un sens à ces données ouvre des perspectives passionnantes. L'acquisition de connaissances à partir des données est l'objectif principal de l'analyse des Big Data. En d'autres termes: il est question de valeur.

Les Big Data génèrent des questions et posent des problèmes non négligeables, en particulier en termes de véracité. Celle-ci se rapporte à la qualité des données, qui peut varier considérablement et nécessite des approches, des règles et des techniques adéquates. Des questions sont également soulevées quant à la protection des données et de la vie privée, exigeant des garanties.

Mais avant de se plonger dans les techniques d'analyse des Big Data, une organisation ou un groupe de parties prenantes intéressés doivent commencer par se demander: quelle est la problématique générale dans notre domaine? Comment nous imaginons-nous la résoudre? Qui a besoin des indications que nous fournirons et qui les utilisera? Quelles seront la portée, la granularité et la visualisation des indications? Qui donnera un sens à ces indications tirées des données?

Les domaines d'application de l'analyse des Big Data sont vastes; heureusement, les phénomènes et la dynamique des marchés de l'emploi et des compétences peuvent être passés au crible et analysés en utilisant les Big Data. Toutefois, un certain nombre de thèmes importants pourraient ne pas encore être pris en compte par l'analyse des Big Data, par exemple les caractéristiques et les tendances de l'emploi informel, qui est très présent dans de nombreux pays.

L'ETF remercie l'équipe d'experts – Mario Mezzanica et Fabio Mercorio – qui ont rédigé ce guide d'introduction, pour avoir partagé leur expérience et leur savoir, en tenant compte de leurs propres

recherches (CRISP, Université de Milan-Bicocca) et d'autres projets pertinents dans le monde, mentionnés à titre d'exemples dans ce guide d'introduction. L'un des cas les plus marquants est le projet en cours «Informations en temps réel concernant le marché du travail et les compétences requises» du Centre européen pour le développement de la formation professionnelle (Cedefop).

Les Big Data pour les SIMT combinent une série d'éléments spécifiques de la transformation numérique, notamment les algorithmes d'apprentissage automatique, l'utilisation de grands volumes de données internet et une architecture informatique spécifique. Ces nouvelles techniques et sources de données vont continuer à évoluer. Et nos compétences et notre compréhension dans ce domaine devront en faire autant. Ce petit guide est une première étape.

L'ETF remercie toutes les organisations qui ont contribué à ce document en fournissant des exemples et des cas qui ont permis d'illustrer les messages clés. L'équipe d'experts, Mario Mezzanzanica et Fabio Mercorio, méritent tout particulièrement d'être salués car ils ont fait preuve de souplesse tout au long du processus de rédaction afin d'adapter les informations aux besoins des utilisateurs cibles. Eduarda Castel-Branco, expert de l'ETF, a coordonné les travaux et les discussions avec les experts et a dirigé le processus de révision au cours duquel de précieux commentaires des experts de l'ETF, Michael Reiner et Martino Rubal Maseda, ont été inclus.

European Training Foundation

Table des matières

Préface 2

Synthèse 6

Partie 1: Big Data et information sur le marché du travail – comment améliorer l'IMT à l'ère numérique. Aperçu, état des lieux, potentiel et limites 10

- Contexte et définitions 10
  - Information/renseignement sur le marché du travail 10
  - Système d'information sur le marché du travail 11
  - Sources de données pour l'IMT 11
  - Questions liées au RGPD en ce qui concerne le MT 14
- Les Big Data au service de l'IMT 15
  - Transformer les Big Data en information sur le MT 18
- Publications sur les Big Data pour l'IMT 26
  - Panorama de l'UE 26
  - Hors de l'UE: projets non européens 26
  - Matrice projets/caractéristiques: un modèle comparatif pour clarifier quels projets ont abordé une préoccupation/un problème spécifique lié aux Big Data pour l'IMT 27
- Les Big Data pour l'IMT en pratique 28
  - Suivi du MT en temps réel Les cas de l'Italie et de l'Espagne 28
  - Le projet Bizkaia Talent 31
  - Le projet ESSnet Big Data 32

Partie 2: Mesures systématiques pour intégrer l'analyse des Big Data dans l'IMT 34

- Composants nécessaires/recommandés d'une architecture Big Data 34
  - Architectures, technologies et outils de pointe 38
  - Le rôle de l'IA pour l'IMT: algorithmes et plates-formes pour l'examen des données du MT 43

Section 3: exemples d'utilisation de l'analyse des Big Data pour les SIMT: sélection de cas à utiliser à titre de références pratiques 50

- CyberSeek.org **[États-Unis d'Amérique]** 50
- WheretheWorkIs.org **[UK]** 51
- Observatoire des talents basques **Espagne:** 53
- La taxonomie des compétences fondée sur des données **[Royaume-Uni]** 54

Technical, Entrepreneurial and Vocational Education and Training (TEVET) Malawi **[Malawi]**. ...55
Projets: (A) Professions de transition et (B) indicateurs de tensions **[Pays-Bas]** 56
Informations en temps réel sur le marché du travail concernant les compétences requises **[tous les États membres]** 57
Section 4: Conclusions et recommandations.....61
Résumé des recommandations et des mesures à l'attention de l'ETF et de ses pays partenaires61
Idées de projets pilotes 63
Les avantages pour les citoyens en termes de parcours professionnels 63
Le rôle des Big Data dans le contexte de l'IMT 63
Le renforcement de l'utilisation des Big Data dans les pays en développement et en transition ...64
Des idées de projets pilotes 64
Références 65

European Training Foundation

SYNTHESE

Au cours des dernières décennies, des forces et des facteurs importants ont radicalement changé la nature et les caractéristiques du marché du travail, tant dans les pays avancés que dans les pays en développement. D'une part, le progrès technique, la mondialisation et la réorganisation du processus de production – avec l'externalisation et la délocalisation – ont radicalement modifié la demande de certaines compétences: de nombreux emplois disparaissent alors que de nouveaux apparaissent. Parmi ceux-ci, certains sont simplement une variante d'emplois existants, tandis que d'autres sont de véritables nouveaux emplois qui n'existaient pas il y a quelques années. **En particulier, le vieillissement de la population dans les économies avancées intensifie le besoin de formation continue** et est susceptible d'affecter la demande structurelle de certaines compétences: la quantité et la qualité de la demande de compétences et de qualifications associées au nouveau marché du travail ont considérablement changé. Non seulement de nouvelles compétences sont nécessaires pour occuper de nouveaux emplois, mais les compétences requises pour les emplois existants ont considérablement changé.

D'autre part, ces dernières années, la quantité d'information sur le marché du travail (IMT) véhiculée par les portails et services internet spécialisés a connu une croissance exponentielle, encourageant et soutenant la réalisation de nombreux services et outils internet liés au marché du travail, tels que les services d'adéquation entre l'offre et la demande d'emplois, la publication de postes à pourvoir, les services de mise en commun de programmes d'enseignement et la création d'un réseau de professionnels qui partagent et échangent librement des opportunités sur le marché du travail.

Dans un tel scénario dynamique, des questions se posent sur la façon d'observer, comprendre et analyser correctement l'impact d'internet sur le marché du travail et sa dynamique, comme les suivantes:

Quels métiers vont se développer à l'avenir et où? Quelles compétences seront les plus demandées par les entreprises dans les prochaines années? Quelles sont les compétences à acquérir lors de son apprentissage tout au long de la vie? Quels sont les emplois réellement nouveaux et ceux qui ne sont qu'une évolution d'anciens emplois existants qui requièrent des compétences nouvelles ou technologiques? Quel est l'impact de la numérisation au sein des professions? Quel rôle jouent les compétences non techniques dans les emplois existants et quelles sont les compétences non techniques les plus importantes à acquérir?

Ce ne sont là que quelques-unes des questions placées au premier plan du débat politique entre les économistes, les décideurs politiques et les experts du marché du travail d'une manière générale. Aujourd'hui, ces questions doivent être abordées en se concentrant sur des paradigmes axés sur les données qui nous permettent d'observer et de suivre un phénomène (i) en temps utile, (ii) de manière inductive, c'est-à-dire à ce que les données soient utilisées pour formuler et confirmer plutôt qu'infirmer des hypothèses et (iii) de façon très précise.

En effet, **l'IMT sur internet offre une grande opportunité de suivi du marché du travail en temps réel**, afin de mieux comprendre la dynamique du marché du travail, saisir les besoins et les tendances en matière de compétences en se concentrant sur différentes dimensions (par exemple, le territoire, les secteurs) à une échelle détaillée, c'est-à-dire les **Big Data relatives au renseignement sur le marché du travail (Big Data pour le renseignement sur le marché du travail)**. Il n'est pas surprenant qu'il y ait un intérêt croissant pour la conception et la mise en œuvre de véritables applications d'IMT pour les données sur le marché du travail présentes sur internet afin de soutenir les activités d'élaboration et d'évaluation des politiques par le biais d'une prise de décision fondée sur des preuves. Ceci représente l'objectif de l'IMT, un domaine qui devient de plus en plus pertinent pour l'élaboration et l'évaluation des politiques du marché du travail de l'Union européenne (UE).

En 2016, la Commission européenne a souligné l'importance des activités professionnelles et éducatives. En effet, elles sont appréciées parce qu'elles «favorisent les compétences spécifiques du poste de travail concerné autant que les compétences transversales, facilitent le passage à l'emploi et permettent le maintien et la mise à jour des compétences de la main-d'œuvre en fonction des besoins sectoriels, régionaux et locaux». En 2016, l'UE et Eurostat ont lancé le projet ESSnet Big Data, impliquant 22 États membres de l'UE dans le but d'«intégrer les Big Data dans la production régulière des statistiques officielles, en utilisant des pilotes qui explorent le potentiel de sources de Big Data sélectionnées et conçoivent des applications concrètes». La même année, le Centre européen pour le développement de la formation professionnelle (Cedefop) a lancé un appel d'offres afin de mettre en place un système d'analyse des postes à pourvoir publiés en ligne et de développer un système ou un outil permettant au Cedefop d'analyser les postes à pourvoir et les nouvelles compétences requises dans tous les États membres de l'UE, en réalisant un système multilingue à part entière (32 langues) qui collecte les postes à pourvoir, extrait les compétences et effectue un suivi en temps réel dans les 28 États membres de l'UE afin de faciliter la prise de décision.

Bien que ces initiatives diffèrent, la référence commune repose sur la reconnaissance de l'énorme pouvoir informatif de l'IMT sur internet. Ce pouvoir informatif peut être exploité en réunissant des informaticiens, des statisticiens, des économistes et des experts du marché du travail pour obtenir des connaissances utiles du marché du travail à partir de données brutes afin de comprendre la dynamique et les tendances du marché du travail sur internet qui s'orientent vers un processus de prise de décision fondée sur les données grâce à l'IMT.

Ce rapport examinera les avantages, le potentiel, les limites, les problèmes méthodologiques et techniques, les questions relatives à la recherche, ainsi que les projets et les études de cas réels relatifs à l'utilisation des Big Data pour l'IMT. Nous introduirons le sujet en examinant le rôle des Big Data dans le contexte du marché du travail, et en faisant le point sur l'application dans le domaine de l'IMT. Ensuite, nous discuterons de certains aspects techniques nécessaires à l'intégration de l'analyse des Big Data dans l'IMT. Des exemples ainsi que des applications et des projets récents (à la fois dans l'UE et en dehors de l'UE) seront fournis, exposant les objectifs, les données et les sources utilisées, les résultats obtenus, ainsi que les questions ouvertes et difficiles pour chaque projet rapporté. Enfin, nous

European Training Foundation

formulerons un ensemble de recommandations et de mesures et nous fournirons quelques idées de projets qui ont émergé de la conférence de l'ETF qui s'est tenue à Turin en novembre 2018.

PARTIE 1: BIG DATA ET INFORMATION SUR LE MARCHE DU TRAVAIL – COMMENT AMELIORER L'IMT A L'ERE NUMERIQUE. APERÇU, ETAT DES LIEUX, POTENTIEL ET LIMITES

Contexte et définitions

La présente partie introduit brièvement quelques termes et notions de base relatifs aux données sur le marché du travail (MT) qui facilitent la lecture de ce document.

Information/renseignement sur le marché du travail

Les termes «information» et «renseignement» – souvent utilisés de manière interchangeable – font référence aux données relatives aux phénomènes et aux dynamiques du MT qui sont utiles pour faciliter la prise de décision, l'élaboration et l'évaluation des politiques. En anglais, l'abréviation LMI faisant référence aux deux termes, on ne peut pas savoir clairement s'il est question de l'information ou du renseignement sur le MT.

Plus précisément, le terme «information» décrit toutes sortes de données et d'informations utilisées pour soutenir les activités opérationnelles liées au MT (pas d'analyse) ainsi que toute information relative à l'offre et à la demande du MT. Il s'agit par exemple d'offres de postes à pourvoir, de compétences, de métiers et de CV de candidats à un emploi.

En revanche, le terme «renseignement» est un concept émergent dans l'ensemble de la communauté du MT, en particulier dans l'Union européenne (UE). Bien qu'il n'existe pas de définition unifiée du renseignement sur le MT, il peut être décrit comme la conception et l'utilisation d'algorithmes d'intelligence artificielle (IA) et de cadres pour analyser les données liées au MT (aussi appelées information sur le marché du travail) afin de soutenir les politiques et la prise de décision (voir, par exemple les références [1], [2], [3]).

Q: Quand l'information devient-elle du renseignement?

R: Globalement, les données brutes ou agrégées, y compris celles suivies au fil du temps, utilisées en vue de soutenir les activités opérationnelles, relèvent toujours de l'information. L'information devient du renseignement lorsqu'un algorithme automatisé (aujourd'hui principalement basé sur l'IA) traite les données pour générer des indications utiles à des fins d'analyse (par exemple, des prévisions pour les activités de prise de décision, l'apprentissage automatique pour la classification ou encore l'extraction d'informations pour les compétences sur les CV). Notamment, la capacité de traiter et d'analyser des masses de données en temps réel permet d'utiliser les connaissances obtenues par le processus de renseignement dans des systèmes habituellement consacrés au soutien des activités opérationnelles.

Dans un tel scénario, le renseignement sur le MT doit être considéré comme une activité qui – comme prévu – produit un résultat, appelé «connaissance du MT». Ici, la définition générale de la connaissance s'applique, à savoir, des indications et des informations supplémentaires extraites de l'expérience (dans ce cas, de l'information sur le MT), qui peuvent accroître la prise de conscience et la compréhension du phénomène observé. Cette connaissance permet aux utilisateurs de faire des prévisions et des analyses (comme nous le verrons ultérieurement).

Q: L'information et le renseignement peuvent-ils fonctionner ensemble au sein d'un système (ou d'un cadre) pour soutenir les activités de prise de décision?

R: Oui, c'est la façon dont l'information et le renseignement sur le MT devraient interagir, à savoir au sein d'un système (back-end) qui collecte l'information sur le MT et utilise l'IA pour créer le renseignement sur le MT. Le résultat, la connaissance du MT, est ensuite fourni à un ensemble de parties prenantes en fonction de leurs besoins et de leurs capacités à comprendre la dynamique du marché du travail. **Ce processus décrit comment un système d'information sur le marché du travail (SIMT) doit fonctionner.**

Système d'information sur le marché du travail

Dans la partie précédente, nous avons clarifié la différence entre l'information sur le MT (c'est-à-dire les données brutes normalement utilisées pour échanger des informations dans le cadre des processus des services opérationnels relatifs au MT), et le renseignement sur le MT (outils, algorithmes et procédures pour manipuler l'information sur le MT). Ces deux concepts participent à la réalisation d'un SIMT, où un système d'information (SI) est communément défini comme un ensemble de composantes interconnectées (technologiques et architecturales) qui agissent ensemble pour collecter, récupérer, traiter, stocker et distribuer les informations afin de faciliter les activités telles que la planification, le contrôle, la coordination, l'analyse et la prise de décision dans les organisations professionnelles. **Par conséquent, la valeur de l'information accessible par le biais d'un SI est double: premièrement, elle soutient les activités opérationnelles, et deuxièmement, elle aide les décideurs à atteindre leurs objectifs d'analyse.**

SIMT (intuition)

Un SIMT peut être considéré comme un exemple de système d'information classique qui utilise l'information et le renseignement sur le MT pour soutenir les activités opérationnelles et de prise de décision.

En substance, le concept de SIMT peut être décrit comme un ensemble d'outils capables d'extraire, d'analyser et de diffuser des informations relatives au MT. Néanmoins, il n'existe pas de définition unifiée de ce que devrait être un SIMT, ni de conseils pratiques uniques sur le développement d'un SIMT, car son architecture, ses données et ses méthodes dépendent des besoins d'analyse, qui sont fonction du contexte (par exemple, le pays, l'institution, les questions importantes à traiter, les priorités politiques et l'infrastructure de données). Des exemples pratiques et différents de SIMT figurent dans les récents travaux [4], [5], [6], [2], pour n'en citer que quelques uns. Certains d'entre eux seront abordés ultérieurement dans ce document. À cet égard, la disponibilité des données sur internet (voir partie 2) met en lumière l'importance de mettre à niveau et de faire évoluer les SIMT afin d'inclure les données d'internet et d'exploiter les algorithmes d'IA pour obtenir des indications utiles et formuler des prévisions sur la dynamique et les tendances du MT (comme l'ont récemment exposé Johnson [7] et Frey et Osborne [8] pour prédire le risque de robotisation). Ces raisons ont conduit les analystes et les experts du MT à inclure internet comme source supplémentaire de données et d'information sur le MT dans leur propre travail, afin de mieux décrire et comprendre le MT dans son ensemble.

Sources de données pour l'IMT

Les données administratives, les données statistiques et les données d'internet sont trois grandes catégories de données qui peuvent se combiner pour expliquer un phénomène. Ce très bref aperçu des trois principaux types de données met en évidence leurs particularités et leurs similitudes.

Données administratives: en substance, une définition fiable des données administratives est «l'ensemble des données collectées par les institutions ou organismes gouvernementaux à des fins fiscales, de prestations ou d'administration publique»[9]. Cela signifie que ces données se rapportent également à des informations collectées auprès de (ou sur) des personnes, lesquelles peuvent avoir besoin d'intervenir pour faire partie d'un système qui utilise des données administratives (par exemple, l'enregistrement des agriculteurs dans le système fiscal et de sécurité sociale) ou non (par exemple, le droit du travail italien stipule qu'un système doit automatiquement contrôler le début/la fin de chaque contrat de travail, voir la référence [10]).

Données statistiques: des données statistiques (également appelées données d'enquête) sont collectées pour répondre à un objectif statistique spécifique et prédéfini afin de garantir une couverture donnée de la population, des définitions, de la méthodologie, de la qualité et du temps afin de répondre aux besoins analytiques des parties prenantes (voir, par exemple la référence [11]). Il est clair que l'utilisation de données administratives à des fins statistiques est loin d'être simple, car elle implique des questions difficiles telles que l'identification de la population, des groupes cibles et de la taille de l'échantillon, et la difficulté de sélectionner la variable modèle pour échantillonner la population.

Bien que les données statistiques et administratives diffèrent en termes d'objectifs, elles partagent certaines caractéristiques intéressantes, comme le montre le tableau 1. Premièrement, les données statistiques sont souvent des données structurées (par exemple des tableaux avec des chiffres dont la structure et le type sont bien définis), tandis que les données administratives peuvent également comprendre des données semi-structurées, dont la structure est partiellement définie et où du texte libre peut apparaître. Néanmoins, ces données peuvent être facilement stockées en utilisant des paradigmes relationnels classiques (par exemple, le langage de requête structuré – SQL). La garantie que les données statistiques sont de bonne qualité est la responsabilité du producteur ou du propriétaire des données qui a également élaboré la collecte de données/l'étude. Cela peut ne pas être vrai pour les données administratives, dont la qualité peut être considérée comme suffisante pour le propriétaire des données mais médiocre pour le consommateur des données. Cela n'est pas surprenant, car la qualité des données est définie comme «l'aptitude à l'emploi», et la satisfaction de la qualité peut donc varier selon l'utilisateur. Les données administratives sont collectées pour suivre un phénomène plutôt qu'à des fins d'analyse (voir, par exemple la référence [12]). Cela signifie également que la crédibilité des données statistiques – c'est-à-dire «la mesure dans laquelle les données sont acceptées ou considérées comme vraies, réelles et crédibles» [13] – dépend de la fiabilité du producteur/propriétaire des données, et ceci peut également être vrai pour les données administratives. Comme ces deux types de données sont collectées à partir d'un système (administratif) ou conçues pour un objectif d'analyse spécifique (statistique), leur valeur est intrinsèque. En d'autres termes, les données ont une valeur intrinsèque. Cette valeur demeure, même si elle peut être améliorée par l'analyse et la mise en relation des données.

Ce scénario change lorsqu'il s'agit des **données d'internet**, qui se rapportent simplement à toutes les données provenant de sources publiées sur internet. Comme on pourrait l'imaginer, ces données peuvent **avoir différents types de structure**, **elles peuvent ainsi être structurées** (par exemple des tableaux recueillis sur internet), **semi-structurées** (par exemple au format XML[1], comme les tweets) **ou pas du tout structurées** (toute autre donnée). Ces données sont générées en continu à partir d'une ou plusieurs sources internet, sur lesquelles l'utilisateur n'a aucun contrôle, ce qui l'oblige à surveiller et à collecter des données en permanence. Comme la structure des données d'internet peut varier de manière imprévisible, les paradigmes relationnels (qui nécessitent une structure de données fixe et définie) ne peuvent pas être utilisés pour stocker

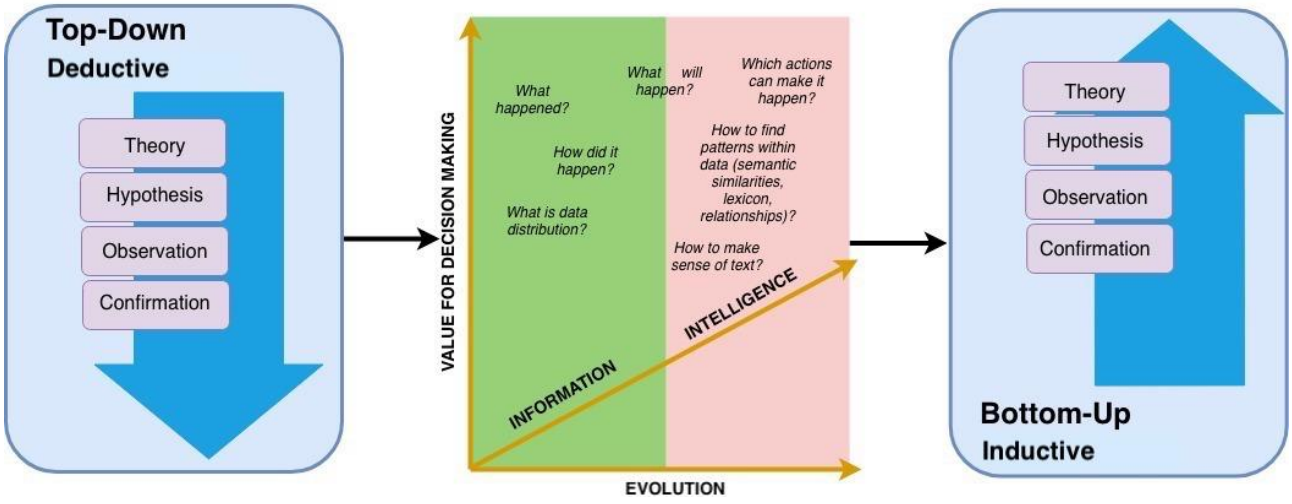
[1] Le **langage de balisage extensible (XML**) est un langage de balisage utilisé pour définir un ensemble de règles pour l'encodage de documents dans un format à la fois lisible par l'homme et par machine.

les données d'internet telles qu'elles affluent sur les sites web. Des paradigmes NoSQL[2] ont été développés pour traiter cette question. En outre, la **qualité dépend de la capacité de l'utilisateur à identifier les problèmes au sein même des données** (duplications, données manquantes, coquilles, synonymes, etc.) **mais aussi de la couverture, qui doit être estimée et mesurée par l'utilisateur des données, en combinant souvent plusieurs sources de données sur internet**. Par conséquent, la **crédibilité dépend de la fiabilité de l'utilisateur des données** plutôt que de celle du propriétaire des données. Enfin, **les données d'internet n'ont pas de valeur intrinsèque. Leur valeur dépend de leur capacité à décrire et à expliquer un phénomène**. **En d'autres termes, les données d'internet sont brutes et leur valeur doit être trouvée/découverte par l'utilisateur.**

Les données d'internet pourraient être assimilées à un bloc de granit devant être travaillé par un sculpteur, qui peut décider d'utiliser un appareil de fraisage ou un burin, pour façonner une figure plutôt qu'une autre.

Cette principale différence par rapport aux données statistiques et administratives oblige également l'utilisateur à modifier son approche de l'analyse, passant d'une approche basée sur un modèle qui utilise un processus descendant à une approche fondée sur les données qui nécessite une méthode ascendante, comme le montre la Figure 1.

Figure 1 Un changement de paradigme, d'une approche descendante à une approche ascendante



[2] Le terme NoSQL (Not only SQL) fait référence à un mouvement grandissant pour faciliter le stockage et la requête de données non structurées. Le rôle du mouvement NoSQL au sein des SIMT est abordé dans la partie 2.

Q: En raison de la présence massive de données/textes non structurés, il semble que l'utilisation des Big Data rende impossible l'exécution des tâches d'évaluation de la qualité des données, qui sont bien définies pour les données structurées. Est-ce le cas?

R: Le débat reste ouvert quant à la nécessité ou non de soumettre les données d'internet à une évaluation de la qualité (et à un nettoyage). Certains pensent que les données d'internet devraient être gérées comme des données structurées classiques, tandis que d'autres affirment que le principe «garbage in, garbage out » (si vous entrez de mauvaises données, vous obtiendrez de mauvais résultats) ne s'applique pas aux Big Data, puisque le volume aura tendance à éliminer le bruit. Selon notre expérience, la qualité des Big Data dépend principalement de la fiabilité des sources utilisées pour collecter les données. Le classement des sources internet est donc crucial. Toutes les techniques d'évaluation de la qualité des données peuvent être appliquées: qu'elles soient fondées sur des règles (si un modèle de données peut être identifié) ou statistiques (identifier les aberrations et supprimer le bruit des données).

Tableau 1 Principales caractéristiques des sources de données sur le MT

Type de source de données sur le MT	*Type de données*[3]	*Taux de génération*	*Paradigme du modèle de données*	*Qualité*	*Couverture*	*Paradigme de l'analyse*	*Crédibilité*	*Valeur*
Statistiques	Structurées	Périodiquement	Relationnel	Responsabilité du propriétaire	Responsabilité du propriétaire	Approche descendante et basée sur un modèle	Responsabilité du propriétaire	Intrinsèque
Administratives	Structurées ou semi-structurées	Périodiquement	Relationnel	Responsabilité du propriétaire	Responsabilité du propriétaire et de l'utilisateur	Approche descendante et basée sur un modèle	Responsabilité du propriétaire et de l'utilisateur	Intrinsèque
Internet	Structurées, semi-structurées ou non structurées	Temps quasi-réel ou temps réel	Relationnel et non relationnel (NoSQL)	Responsabilité de l'utilisateur	Responsabilité de l'utilisateur	Approche ascendante et fondée sur les données	Responsabilité de l'utilisateur	Extrinsèque

Questions liées au RGPD en ce qui concerne le MT

Le règlement général sur la protection des données (RGPD) est entré en vigueur en mai 2018 dans tous les États membres de l'UE. Il représente un premier pas vers une réglementation de la manipulation et du traitement des **données à caractère personnel**.

Lorsque les données ne contiennent pas d'informations personnelles, le RGPD ne s'applique pas. Lorsque, par contre, les données contiennent des informations personnelles liées à un sujet (par exemple des données relatives au CV, aux préférences personnelles, à l'historique du parcours professionnel ou aux compétences personnelles), alors le SIMT qui utilise les données doit être conforme au RGPD.

3 Les données structurées font référence à des types de données clairement définis dont la structure et le schéma récurrent les rendent facilement consultables par un système automatisé. Les données non structurées font référence aux données dont la structure ne peut pas être facilement définie comme un schéma ou un type, ce qui rend la recherche dans ces données difficile (par exemple, les textes libres, les fichiers audio, vidéos et les messages des médias sociaux). Les données semi-structurées désignent les données dont la structure est partiellement définie (par exemple, les documents XML).

En substance, le RGPD vise à garantir les droits fondamentaux de la personne concernée et à accroître la responsabilité des entreprises qui contrôlent et traitent les données à caractère personnel. Le RGPD établit un certain nombre de restrictions et de contraintes sur l'utilisation des données à caractère personnel:

- Premièrement, le droit de la personne concernée d'accéder aux informations collectées à son sujet impose des restrictions à la prise de décision automatisée par les entreprises et les organisations qui utilisent ces données.

- Deuxièmement, les entités désignées pour effectuer le traitement des données à caractère personnel (c'est-à-dire les sous-traitants) doivent informer les personnes concernées des données collectées (articles 13 à 15).

- Troisièmement, la transparence joue un rôle clé, obligeant le sous-traitant à traiter les données de manière transparente (article 5, paragraphe 1, point a), par un traitement transparent des données (article 13, paragraphe 2 et article 14, paragraphe 2), et à notifier à la personne concernée si une décision individuelle automatisée est appliquée à ses données à caractère personnel (article 22). En outre, les articles 13 et 14 stipulent que, lorsqu'un profilage a lieu, la personne concernée a le droit d'obtenir «des informations utiles concernant la logique sous-jacente».

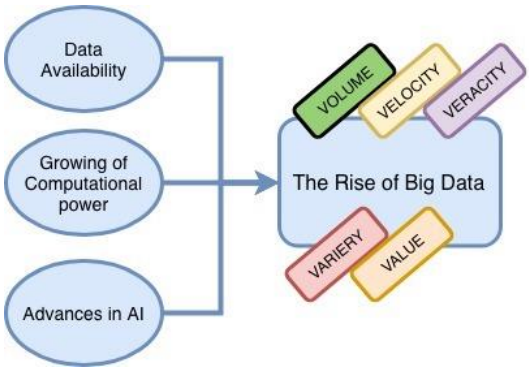
D'un point de vue technique, ces règles s'appliquent également à tous les processus et routines d'extraction, de transformation et de chargement (ETC[4]), qui extraient des données, les transforment d'un format à un autre et, enfin, chargent les données traitées et agrégées dans des entrepôts de données à des fins d'analyse. Ainsi, lorsqu'un indicateur clé de performance ou un indicateur général d'activité est calculé, il n'est pas possible d'identifier la source d'information ou les données relatives à une personne concernée qui ont été utilisées. Cela est également valable pour les données (personnelles) relatives au MT. Globalement, cela signifie que le sous-traitant est chargé de garantir, entre autres: (i) que la personne à laquelle se réfèrent les données ne peut être identifiée, directement ou indirectement, lorsque l'identifiant de la personne concernée est un nom, un numéro d'identification, des données de localisation, un identifiant en ligne ou un ou plusieurs éléments caractéristiques de son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale (article 4); (ii) que les données sont traitées de manière licite, loyale et transparente au regard de la personne concernée (article 5); et (iii) que les données sont collectées pour des finalités déterminées, explicites et légitimes et ne sont pas traitées ultérieurement d'une manière incompatible avec ces finalités (article 5).

Les Big Data au service de l'IMT

L'intérêt croissant suscité par la manipulation et l'utilisation des Big Data a permis aux gestionnaires de mesurer plus efficacement leurs activités, permettant ainsi d'acquérir des connaissances qui peuvent améliorer leur prise de décision et leurs performances (voir, par exemple la référence [14]). Il s'agit d'une remarque très générale qui s'applique également au MT. Cependant, ce que sont réellement les Big Data, la véritable nature des Big Data, leur importance et leurs limites, ainsi que les défis et les opportunités associés au traitement des Big Data, sont autant de questions qui restent ouvertes au débat.

[4] L'ETC est une approche qui soutient les tâches de prétraitement et de transformation des données dans le processus d'extraction de connaissances à partir de bases de données (ECD). Les données extraites d'un système source subissent une série de transformations (analyse, manipulation, nettoyage) avant le chargement dans un entrepôt de données.

Figure 2 Éléments clés de l'essor des Big Data: (i) disponibilité des données, (ii) puissance de calcul toujours plus grande et (iii) récents progrès en matière d'IA



Ces dernières années, la communauté a essayé de répondre à ces questions en utilisant une variété de «modèles» de Big Data, basés sur les dimensions (les cinq «V») qu'une application/approche des Big Data devrait posséder. Bien que plusieurs modèles aient été proposés, nous suggérons ici un modèle des cinq V adapté au domaine du MT, qui caractérise les Big Data selon cinq dimensions fondamentales:

Volume En 2017, il y avait environ 4 milliards d'internautes dans le monde. Ce nombre augmente à un rythme rapide: le premier milliard a été atteint en 2005, le deuxième milliard en 2010 et le troisième milliard en 2014. Environ 40 % de la population a accès à une connexion internet. En 2018, on comptabilisait environ 2 milliards de sites web actifs (sans compter le «web profond», c'est-à-dire les pages web qui ne peuvent être indexées par les moteurs de recherche) et plus de 3,5 milliards de recherches effectuées sur Google chaque minute[5]. Chaque seconde, plus de données traversent internet que celles qui étaient stockées sur l'ensemble du réseau internet il y a seulement 20 ans. Cela donne aux entreprises une occasion unique d'accéder à ces données et de les collecter afin de prendre de meilleures décisions et d'améliorer leurs activités. On estime, par exemple, que Walmart est capable de collecter environ 2,5 pétaoctets (soit 2,5 quadrillions d'octets) de données par heure à partir des transactions de ses clients. Bien que l'approche des Big Data classique mesure le volume en octets, ce qui fonctionne bien pour les données utilisateur générées par un système (par exemple, les historiques et les transactions), cette unité de mesure ne s'applique pas à l'information sur le MT, car l'échelle change considérablement. Dans le domaine du MT, il pourrait être intéressant de mesurer le nombre de documents ou d'éléments collectés qui se rapportent à la demande ou à l'offre du MT, ou le nombre de sources relatives au MT considérées.

Vélocité Cette dimension fait référence au rythme auquel les données sont générées, ou collectées dans le cas de l'IMT. Ces données sont collectées auprès de sources tierces qui peuvent décider d'autoriser la collecte autonome de données par (i) des API[6], (ii) des procédures par lots exécutées périodiquement ou (iii) par «crawling» [7] ou «scraping» en temps quasi réel de la source et l'obtention de données à intervalles rapprochés et fixes. Il est clair que plus la fréquence de la collecte de données est faible, plus le volume de données

[5] Source: Internet Live Stats: http://www.Internetlivestats.com/.
[6] API (Application Programming Interface) désigne une interface de programme d'application. Dans le contexte du web, il s'agit d'un ensemble de procédures qu'un utilisateur utilise pour communiquer avec le serveur (par exemple pour la collecte de données). Par conséquent, la communication entre l'utilisateur et le serveur est réglementée, bien définie et contrôlée.
[7] Le crawling (exploration) collecte les données des sites web tels qu'ils sont, tandis que le scraping (récupération) identifie certaines parties d'un site web à collecter. En substance, le scraping analyse et identifie les données que l'utilisateur souhaite collecter, alors que le crawling collecte simplement tout le contenu web.

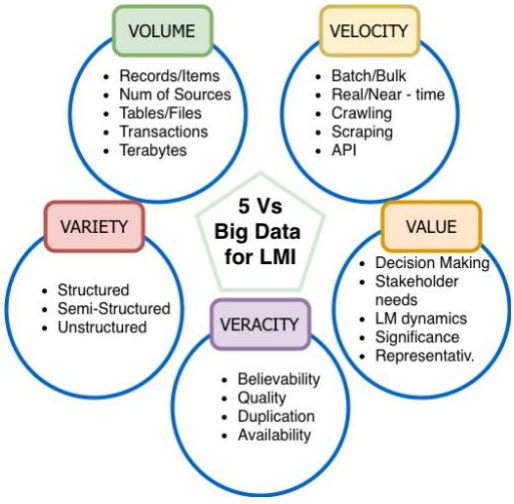
collectées est important et plus on a besoin de disposer de ressources informatiques et de stockage importantes.

Variété Cette dimension fait référence à la variété des types de données au sein des sources de Big Data, comme indiqué dans le tableau 1. La source peut être structurée, semi-structurée ou complètement non-structurée. En raison de l'utilisation de grands volumes de contenu non structuré, les lexiques utilisés dans chaque source varient, et l'utilisation à grande échelle du langage naturel signifie que l'hétérogénéité des données est considérable.

Véracité La véracité des données indique le degré d'exactitude ou de confiance d'un ensemble de données. Comme nous l'avons vu précédemment, la qualité des données d'internet ne peut être manipulée à la source, mais doit être évaluée lors de la collecte et du stockage des données, au moyen d'analyses, de procédures et d'outils ad hoc. Les biais, les anomalies ou les incohérences, les doublons et la volatilité sont quelques-uns des aspects qui doivent être supprimés pour améliorer l'exactitude des Big Data. Comme on peut l'imaginer, pour une source de données considérée, plus la variété est élevée, plus la véracité est élevée. En effet, l'utilisation du langage naturel apporte beaucoup de bruit ne contenant aucune information dans un texte (par exemple les prépositions, les termes sans rapport avec le sujet considéré, les conjonctions et les acronymes qui doivent être développés). Toutes ces questions doivent être correctement traitées pour permettre aux données non structurées de produire des connaissances dans les étapes du processus d'extraction de connaissances à partir de bases de données (ECD).

Valeur Enfin, les données doivent présenter un intérêt pour un domaine ou un objectif spécifique. En d'autres termes, comme nous l'avons vu précédemment, les données d'internet n'ont pas de valeur intrinsèque. Leur valeur réside dans les connaissances que l'utilisateur en extrait pour expliquer un phénomène, ou pour faciliter la prise de décision par des analyses et des recommandations. Il convient de mentionner l'importance de l'analyse des besoins des parties prenantes, qui doit permettre de déterminer quelle composante des connaissances intéresse une partie prenante donnée et laquelle ne l'intéresse pas. Dans le cas des postes à pourvoir publiés sur internet, il pourrait être intéressant pour un utilisateur à la recherche d'un nouvel emploi d'analyser ses lacunes en matière de compétences, tandis qu'un analyste du MT pourrait s'intéresser à observer le MT sur internet dans son ensemble à un niveau régional spécifique. Les mêmes connaissances peuvent avoir des points d'accès différents selon les besoins des parties prenantes.

Figure 3 Modèle de Big Data des cinq V, adapté à l'IMT



Transformer les Big Data en information sur le MT

L'extraction des connaissances à partir des (méga)données sur le MT a été examinée. À cette fin, la Figure 5 Processus ECD et dimensions V des Big Data impliquées dans chaque étape Figure basée sur une figure de la référence [16]

résume les principaux défis liés à la gestion de l'IMT sur internet, tels qu'ils sont abordés dans le document [15]. Une approche qui permet de gérer les Big Data pour l'IMT est le processus ECD. Le processus ECD comprend cinq étapes principales, comme le montre la référence [16] dans la Figure 5. Il est évident qu'il faut l'adapter au domaine considéré, en valorisant une tâche ou une étape par rapport à une autre.

Figure 4 Vue schématique des principaux éléments (i) du scénario du MT sur internet, (ii) de certains besoins des parties prenantes et (iii) des actions que nous proposons

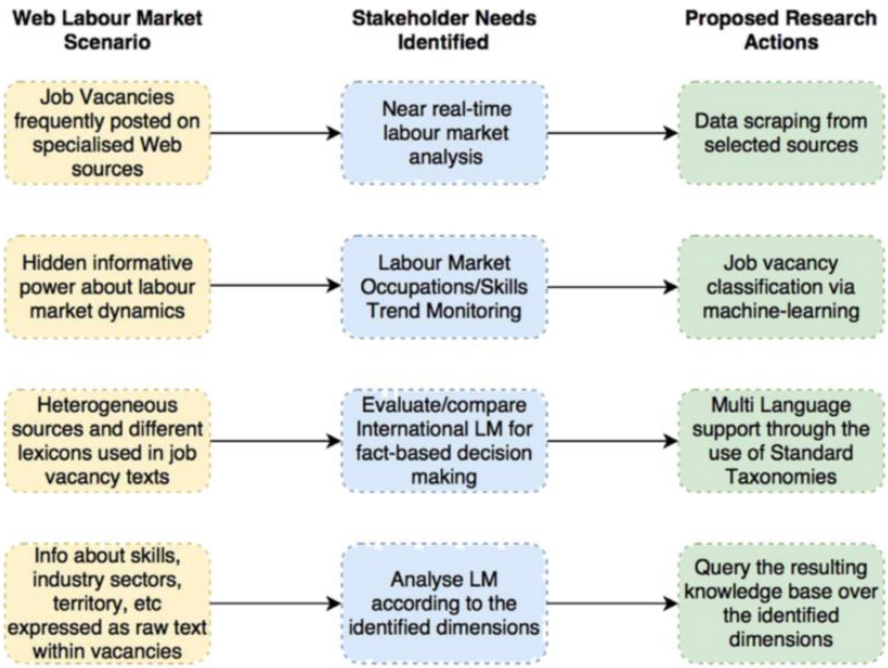
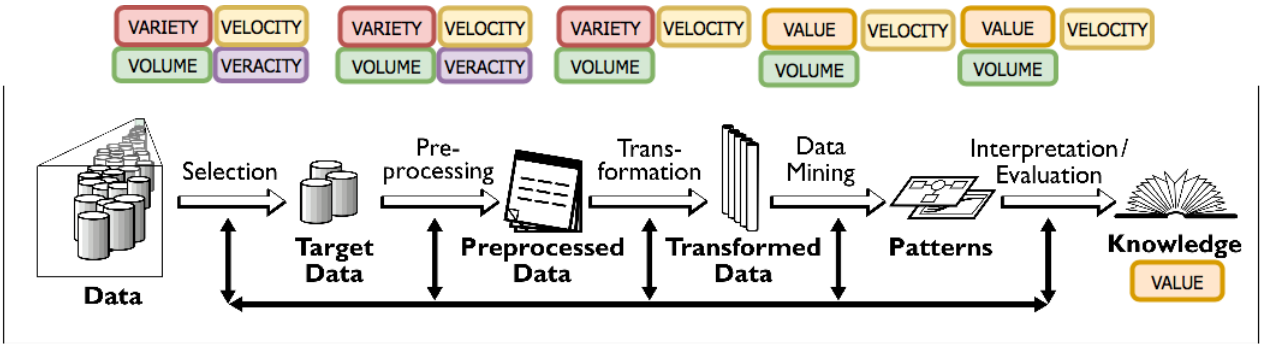


Figure 5 Processus ECD et dimensions V des Big Data impliquées dans chaque étape Figure basée sur une figure de la référence [16]



Étape 1: Sélection La sélection des sources de données est la première étape. Chaque source internet doit être évaluée et classée en fonction de la fiabilité des informations. Par exemple, cette phase doit tenir compte

de la date de publication de l'offre d'emploi, de la fréquence de mise à jour du site web, de la présence de données structurées et de toute restriction de téléchargement. À la fin de cette phase, un classement des sources internet fiables est établi. Cette étape concerne les cinq dimensions V des Big Data, y compris la véracité, c'est-à-dire les biais, le bruit et les anomalies présentes dans les données. Les questions clés que la phase de sélection a soulevées auprès des experts du MT sont:

1. [Question d'ordre statistique] Comment identifier les critères à inclure dans le modèle de source et comment extraire ces critères (c'est-à-dire les variables) des sources? Comment classer les sources?

2. [Question d'ordre technique] Comment identifier un paradigme de modélisation de données (par exemple relationnel, document, valeur clé, graphique) pour stocker une quantité considérable de données à grande échelle? Comment collecter automatiquement les données? Avons-nous besoin d'un accès à une API ou devons-nous développer un scraper/crawler? Comment programmer les processus de collecte automatique des données?

3. [Expert dans le domaine du MT] Comment sélectionner les bonnes sources? Avons-nous sélectionné les bonnes sources?

Étape 2: Prétraitement Cette étape consiste à nettoyer les données pour éliminer le bruit ou les aberrations (le cas échéant), à décider de la manière de traiter les données manquantes et à identifier une fonction pour détecter et supprimer les doublons d'entrées (par exemple, les doublons d'offres d'emploi ou les offres d'emploi présentant des valeurs manquantes). L'évaluation de la qualité et le nettoyage des données sont des tâches essentielles dans toute approche de prise de décision fondée sur les données, afin de garantir la crédibilité du processus global, c'est-à-dire «la mesure dans laquelle les données sont acceptées ou considérées comme vraies, réelles et crédibles» (voir, par exemple les références [12], [13], [17]).

L'identification des doublons de postes à pourvoir est loin d'être simple. Les postes à pourvoir sont généralement publiés sur plusieurs sites web, ce qui constitue une duplication, alors que la réutilisation du même texte pour annoncer un poste similaire ne l'est pas. L'identification des caractéristiques appropriées pour une reconnaissance correcte des doublons est cruciale dans le domaine du MT sur internet. L'étape de prétraitement réduit la complexité du scénario des Big Data, en atténuant l'impact de la dimension de véracité par l'évaluation de la qualité et le nettoyage des données. Les questions clés que l'étape 2 a soulevées auprès des experts du MT sont:

1. [Question d'ordre statistique] Comment évaluer la cohérence des données? Comment mesurer l'exactitude des données? Comment estimer l'importance des données?

2. [Question d'ordre technique] Comment identifier les doublons dans les enregistrements de données? Comment identifier les valeurs manquantes?

3. [Expert dans le domaine du MT] Comment identifier les synonymes du domaine du MT qui contribuent à améliorer l'exactitude des données? Comment identifier les critères qui caractérisent les valeurs manquantes et les doublons?

Étape 3: Transformation Cette étape comprend la réduction et la projection des données, qui visent à identifier un modèle unifié pour représenter les données, en fonction de l'objectif de l'exercice. En outre, elle peut inclure l'utilisation de méthodes de réduction de la dimensionnalité ou de transformation pour réduire le nombre effectif de variables ou pour trouver des représentations invariantes des données. Comme l'étape 2, l'étape de transformation réduit la complexité de l'ensemble des données en prenant en compte la dimension de variété. Elle est généralement réalisée au moyen de techniques ETC, qui permettent de prendre en charge les phases de prétraitement et de transformation des données dans le processus ECD. Globalement, grâce à l'ETC, les données extraites d'un système source subissent une série de procédures de transformation qui les

analysent, les manipulent puis les nettoient avant de les charger dans une base de connaissances. À l'issue de cette étape, dont le résultat est un modèle de données propres et bien définies, le problème de la variété des Big Data devrait être réglé. Les questions clés que la phase de transformation a soulevées auprès des experts du MT sont:

1. [Question d'ordre statistique] Comment mesurer l'exhaustivité du modèle cible identifié? Le modèle cible conserve-t-il l'importance des données à la fin du processus ETC?

2. [Question d'ordre technique] Comment développer des procédures relatives aux Big Data pour transformer les données brutes en un modèle cible de manière évolutive?

3. [Expert dans le domaine du MT] Comment identifier le format des données de destination et la taxonomie[8]?

Étape 4: Fouille de données et apprentissage automatique L'objectif de la présente étape est d'identifier les algorithmes d'IA appropriés (par exemple, pour la classification, la prédiction, la régression, le regroupement, le filtrage des informations) en recherchant des tendances intéressantes sous une forme représentative particulière, en fonction de l'objectif de l'analyse. Plus précisément, dans le contexte de l'IMT, cette étape nécessite généralement d'utiliser des algorithmes dédiés à la classification du texte (par exemple, basés sur l'ontologie ou l'apprentissage automatique) pour construire une fonction de classification permettant de recenser les données dans l'une de plusieurs classes prédéfinies. Cette étape est cruciale car elle est principalement consacrée à l'extraction de connaissances à partir des données. Les questions clés que la phase de fouille des données et d'apprentissage automatique a soulevées auprès des experts du MT sont:

1. [Question d'ordre statistique et technique] Comment sélectionner le meilleur algorithme? Comment régler les paramètres des algorithmes? Comment évaluer l'efficacité des algorithmes? Comment les mettre en œuvre à grande échelle?

2. [Expert du domaine MT] Quelles connaissances doivent être sélectionnées et lesquelles faut-il écarter? Quelle est l'importance des connaissances obtenues pour le MT? Quelles nouvelles indications ont été découvertes grâce à l'IMT? Comment expliquer les résultats du processus de fouille dans une perspective axée sur le MT?

Étape 5: Interprétation/évaluation Cette dernière étape utilise des paradigmes visuels pour représenter visuellement les connaissances obtenues, en fonction des objectifs de l'utilisateur. Dans le contexte de l'IMT, cela signifie qu'il faut tenir compte de la capacité de l'utilisateur à comprendre les données et leur objectif principal dans le domaine de l'IMT. Par exemple, des agences gouvernementales pourraient être intéressées par l'identification des métiers les plus prisés dans leur région; des entreprises pourraient se concentrer sur le suivi des tendances en matière de compétences et sur l'identification de nouvelles compétences pour certains métiers afin de pouvoir concevoir des parcours de formation pour leurs employés. Ces dernières années, de nombreux travaux se sont concentrés sur la production de bibliothèques visuelles prêtes à l'emploi qui mettent en œuvre différents paradigmes narratifs et visuels. Un exemple puissant est D3.js [18], une bibliothèque réactive fondée sur les données pour produire une visualisation dynamique et interactive des données, même dans le contexte des Big Data (voir, par exemple la référence [19]). Les questions clés que la phase d'interprétation/évaluation a soulevées auprès des experts du MT sont:

[8] Le MT est caractérisé par plusieurs taxonomies normalisées, telles que ISCO/O*NET/CTP pour les métiers, ESCO pour les compétences et NACE pour la classification des activités économiques.

1. [Question d'ordre statistique et technique] Comment sélectionner le paradigme de visualisation? Comment sélectionner un modèle de visualisation approprié pour les connaissances que nous voulons visualiser?

2. [Expert du domaine MT] Comment fournir des connaissances appropriées en fonction des besoins des parties prenantes? Comment identifier des chemins de navigation visuels pour chaque partie prenante? Comment récupérer les retours des utilisateurs du MT (le cas échéant)? **Comment mettre la connaissance du MT au service des entreprises?**

Comme on peut le constater, le nombre de questions techniques et statistiques diminue au fur et à mesure de l'état d'avancement du processus ECD, tandis que le nombre de questions et de défis auxquels les experts du MT sont confrontés augmente. En fait, alors que les spécialistes techniques ont la responsabilité principale de traiter quatre des dimensions V des Big Data, il appartient aux experts du MT de traiter la cinquième dimension V des Big Data, la valeur.

Pour clarifier ce point, nous indiquons dans le tableau 2 ci-dessous quelques éléments du MT qui peuvent être analysés, ainsi que les sources d'information et les principaux avantages/problèmes liés à l'analyse des sources.

Tableau 2 Options d'analyse du MT

Éléments du MT	Source	Avantage	Problème
A. Métiers et compétences - Demande - Offre	- Offres d'emploi - Programmes d'enseignement	- Informations précises - Focalisation sur les seules informations pertinentes - Lexique spécifique - Focalisation sur les termes utilisés par le marché plutôt que sur les taxonomies - Identification des métiers similaires sur la base des compétences requises	- Collecte des données - Distinction entre les termes/concepts et le bruit - Compréhension du lexique et des termes spécifiques au domaine - Utilisation de systèmes de classification normalisés pour réduire la complexité et permettre une comparaison transfrontalière - Gestion du multilinguisme
B. Compétences futures et nouveaux métiers émergents	- Offres d'emploi - Programmes d'enseignement	- Focalisation sur les attentes du MT - Identification des tendances et dynamiques futures/actuelles - Analyse des écarts entre les candidats et les nouvelles professions/compétences pour concevoir des parcours d'apprentissage - Identification des métiers similaires sur la base des compétences requises	- Entraînement d'un algorithme pour comprendre quand un terme est une compétence - Formalisation de la signification d'un nouveau métier (combien de fois et pendant combien de temps une offre d'emploi doit-elle être publiée pour devenir une nouvelle profession?)

C. Niveau de compétences numériques par métier	- Offres d'emploi	- Compréhension de la prévalence des compétences numériques dans les métiers - Planification des politiques et des parcours d'apprentissage pour combler les lacunes par rapport aux attentes du MT	- Entraînement d'un algorithme pour comprendre ce qu'est une compétence numérique - Niveaux de compétences informatiques (y compris les compétences non numériques non techniques/techniques)
D. Compétences transversales (également appelées non techniques) (valables et requises pour de nombreux métiers)	- Offres d'emploi - Programmes d'enseignement	- Compréhension de l'impact des compétences transversales au sein du MT - Conception et planification des parcours d'apprentissage pour acquérir des compétences non techniques	- Formalisation de la signification des compétences transversales - Entraînement d'algorithmes pour reconnaître les compétences transversales sous toutes leurs formes, dont la formulation peut varier considérablement dans le langage naturel (par exemple, résolution de problèmes, capacité à résoudre des problèmes)
E. Inadéquation des compétences	- Offres d'emploi - Programmes d'enseignement - Enquêtes	- Capacité à identifier la surqualification/sous-qualification et l'obsolescence des compétences - Amélioration des politiques du MT en fonction des tendances et de la dynamique du MT - Soutien à la conception de parcours éducatifs et professionnels	- Collecte des données - Nettoyage et intégration des données - Conception et sélection d'un modèle d'analyse intégré - Identification des indicateurs d'inadéquation à différents niveaux de précision

Q: Pouvoir prédictif des Big Data: comment les Big Data peuvent-elles contribuer à l'anticipation des compétences (à court et moyen terme)? Quels sont les aspects méthodologiques à prendre en compte, et quels devraient être la portée et le degré d'approfondissement de l'analyse? Avez-vous des exemples de la manière dont l'anticipation des compétences a été améliorée par l'ajout d'indications issues de Big Data?

R: Les Big Data sont cruciales pour l'anticipation des compétences pour deux raisons. Premièrement, les Big Data, telles que celles dérivées des postes à pourvoir publiés sur internet, sont les seules sources d'informations détaillées sur les compétences comme outil alternatif aux enquêtes sur les compétences, qui ne contiennent qu'un ensemble limité de compétences pouvant être évaluées. Deuxièmement, les compétences demandées peuvent varier selon les métiers, les secteurs et les régions. Afin de suivre ces variations, il est nécessaire de disposer de données très granulaires et détaillées qui ne peuvent être obtenues que par le biais des Big Data.

Sur le plan méthodologique, plusieurs défis doivent être relevés. Tout d'abord, les données doivent être liées au métier/au secteur/à la région. Deuxièmement, les compétences doivent être classées selon une taxonomie significative qui peut être utilisée pour l'analyse. Troisièmement, une série chronologique cohérente est nécessaire pour prévoir l'évolution des compétences dans le temps.

Il existe plusieurs exemples de l'utilisation des Big Data pour l'anticipation des compétences. Aux États-Unis, les compétences extraites des postes à pourvoir publiés sur internet sont utilisées par les universités pour anticiper les tendances du MT et adapter les programmes d'enseignement aux besoins du MT.

Q: Comment peut-on coopérer avec succès avec les organismes statistiques et les organes étatiques qui possèdent des données/registres? Comment peut-on surmonter le scepticisme et les inquiétudes quant au caractère innovant des Big Data? Existe-t-il de bons exemples d'une telle coopération?

R: Selon notre expérience, l'utilisation des Big Data n'empêche pas la pertinence ni l'importance de l'utilisation de statistiques et d'enquêtes officielles dans l'analyse de la dynamique et des tendances du MT pour la prise de décision. Pour donner un exemple, l'une des principales critiques liées à l'utilisation des Big Data est qu'elle repose sur l'importance statistique, car les Big Data doivent être utilisées conjointement et mises en correspondance avec des sources statistiques externes (par exemple, des enquêtes sur la main-d'œuvre) dans le but d'améliorer la capacité d'observer plus largement la dynamique du MT.

En outre, il faut considérer que les catégories d'usagers d'internet sont variables et partiellement observables par construction, ce qui rend difficile l'identification d'un échantillon stable qui soit représentatif de l'ensemble des usagers d'internet. Ainsi, on peut utiliser les statistiques officielles pour estimer la pertinence de chaque classe sur les données collectées sur internet afin que les classes surreprésentées (ou sous-représentées) puissent être pondérées en conséquence.

Il convient de noter que l'utilisation des données d'internet ne peut être négligée, car celles-ci vont augmenter dans un avenir proche. Ainsi, le pouvoir informatif que ces données peuvent fournir pour analyser, observer et mesurer un phénomène peut conférer un avantage concurrentiel pour prendre des décisions rapides et fondées sur les données. Un article récent examinant certaines initiatives liées à l'utilisation des Big Data dans de nombreux projets concrets a été publié par Bergamaschi et al. (2016).

Q: Qualité des données: nous savons qu'il s'agit d'un élément essentiel de la science des données. Cependant, quels sont les principaux aspects et dimensions qui doivent être examinés et sérieusement pris en considération pour un projet utilisant des Big Data? Quelles sont les caractéristiques fondamentales de la qualité des données dans les Big Data?

R: Comme indiqué dans la partie 1, l'évaluation de la qualité et le nettoyage des données sont des tâches cruciales et obligatoires pour garantir la crédibilité de tout processus de prise de décision fondée sur les données. Il n'est pas surprenant que la dimension de véracité des Big Data fasse clairement référence aux problèmes de qualité les plus importants même en présence d'une quantité considérable de données. Il n'existe pas encore de liste officielle et fiable des problèmes de qualité des données les plus pertinents pour une application des Big Data, et le débat reste ouvert dans les milieux universitaires. En effet, certains chercheurs pensent que le volume élevé des données est suffisant pour réduire les effets (éventuels) d'une mauvaise qualité des données; d'autres pensent qu'une approche stricte et formelle de la qualité des données (qui évalue notamment la cohérence, l'exhaustivité, l'exactitude et le bien fondé) doit être appliquée même dans le cas d'une quantité considérable de données. La première approche ignore le fait que les données de mauvaise qualité sont répandues à la fois dans les grandes bases de données et sur le web; la seconde ne tient pas compte des coûts (en termes de temps et de puissance de calcul) pour évaluer la qualité et nettoyer les données dans de vastes ensembles de données. Cela dit, nous pensons que les deux approches devraient être appliquées à des fins d'analyse.

Pour donner un exemple, si on se concentre sur la collecte des postes à pourvoir à partir de sources publiées sur internet, il convient d'utiliser une approche stricte et rigoureuse pour classer les sources sur la base de caractéristiques ou de variables des sources, évaluant ainsi la cohérence, l'exactitude et l'exhaustivité (au moins) des variables des sources internet. Cela pourrait inclure, sans s'y limiter les éléments suivants:

- Typologie: désigne la typologie de la source, qui peut être des agences de recrutement, des journaux nationaux, des sites web spécialisés, ou des sites web publics, sectoriels ou d'entreprise;
- Taille: désigne le nombre de postes à pourvoir publiés sur le site web au moment de l'analyse;
- Temps de mise à jour: désigne la fréquence à laquelle le propriétaire de la source web fournit des données nouvelles et actualisées, ainsi que la présence d'horodatages pour marquer exactement le moment où un poste à pourvoir a été publié;
- Qualité de la description: indique dans quelle mesure la page détaillée des postes à pourvoir est normalisée et complète.

En revanche, l'identification des postes à pourvoir «actifs» fait également référence à la dimension de qualité des données (l'exactitude, dans ce cas). Toutefois, lorsqu'un poste à pourvoir a été publié il y a plusieurs semaines ou mois, il n'y a pas de champ dans une publication de poste à pourvoir qui garantit que l'offre est toujours valable (c'est-à-dire que le poste n'a pas encore été pourvu). Dans ce cas, une approche classique ne fonctionnerait pas car ces informations ne sont pas accessibles de manière automatisée et évolutive (des millions d'éléments à traiter à la fois). Par contre, l'inclusion d'un poste à pourvoir dans l'analyse, quelle que soit sa validité, pourrait avoir des effets imprévisibles sur la crédibilité des analyses. Pour ces raisons, il convient d'élaborer un modèle statistique permettant de déduire la date de validité d'un tel poste à pourvoir sur la base des données historiques relatives à des postes similaires (par exemple, en tenant compte de sources similaires, de postes similaires publiés, de l'entreprise qui publie les postes à pourvoir). Un tel modèle garantirait que les dates de validité ont été évaluées en fonction des caractéristiques de l'ensemble des données.

Nous renvoyons le lecteur à Saha, Barna et Divesh Srivastava. «Data quality: The other face of big data», *30ème Conférence internationale sur l'ingénierie des données de l'IEEE, 2014* IEEE, 2014. et à Sadiq, Shazia et Paolo Papotti. «Big data quality-whose problem is it?», *32ème Conférence internationale sur l'ingénierie des données ((CIED) de l'IEEE, 2016* IEEE, 2016, deux publications récentes qui examinent et proposent des orientations pour traiter de la qualité des données dans des applications impliquant des Big Data d'un point de vue méthodologique.

Publications sur les Big Data pour l'IMT

Au cours des dernières années, plusieurs facteurs ont considérablement modifié la nature et les caractéristiques du MT, tant dans les pays industrialisés que dans les pays en développement. Le progrès technique, la mondialisation et la réorganisation des processus de production ont radicalement modifié la demande de certaines compétences, intensifiant le besoin de formation continue, en particulier dans les emplois hautement spécialisés ou fortement touchés par la numérisation. Dans le même temps, la disponibilité des données augmente grâce à la numérisation, à la diffusion des services web (y compris les services liés au MT) et à la disponibilité d'un grand nombre de solutions techniques pour traiter les Big Data. Dans ce contexte, le nombre d'appels à projets liés au MT augmente rapidement. Nous présentons ci-après quelques exemples d'initiatives (européennes et non européennes) liées aux Big Data et à l'information/au renseignement sur le MT.

Panorama de l'UE

Il existe un intérêt croissant pour la conception et la mise en œuvre de véritables applications d'IMT pour les données sur le MT présentes sur internet afin de soutenir l'élaboration et l'évaluation des politiques par le biais d'une prise de décision fondée sur des preuves. En 2010, la Commission européenne a publié une communication intitulée «Donner un nouvel élan à la coopération européenne en matière d'enseignement et de formation professionnels pour appuyer la stratégie Europe 2020» [20] qui vise à promouvoir les systèmes d'éducation en général, et l'enseignement et la formation professionnels (EFP) en particulier. En 2016, la Commission européenne a souligné l'importance des activités professionnelles et éducatives, car elles sont appréciées «pour favoriser les compétences professionnelles spécifiques et transversales, faciliter la transition vers l'emploi et maintenir et actualiser les compétences de la main-d'œuvre en fonction des besoins sectoriels, régionaux et locaux» [21]. En 2016, l'UE et Eurostat ont lancé le projet ESSnet Big Data [22], impliquant 22 États membres de l'UE dans le but d'«intégrer les Big Data dans la production régulière des statistiques officielles, en utilisant des pilotes qui explorent le potentiel de sources de Big Data sélectionnées et conçoivent des applications concrètes». En outre, en 2014, le Centre européen pour le développement de la formation professionnelle (Cedefop) – créé pour soutenir le développement de l'EFP européen – a lancé un appel d'offres pour la mise en œuvre d'une étude de faisabilité et d'un prototype fonctionnel pour collecter et classer les postes à pourvoir publiés sur internet de cinq pays de l'UE [23]. La justification de ce projet est de transformer les données extraites des postes à pourvoir publiés sur internet en connaissances (et donc en valeur) pour l'élaboration et l'évaluation des politiques par le biais d'une prise de décision fondée sur des faits. Vu le succès du prototype, un autre appel d'offres a été lancé pour la création d'un suivi du MT sur internet dans toute l'UE, y compris les 28 États membres et les 24 langues de l'UE [24].

Il convient également de mentionner le projet LMI4All [25], un portail de données en ligne qui connecte et normalise les sources existantes d'IMT fiables et de qualité, dans le but d'éclairer les décisions de carrière. Les données sont mises à disposition gratuitement via des API pour être utilisées sur des sites web et des applications tierces.

Hors de l'UE: projets non européens

Un problème critique concernant le MT est la possibilité que des personnes perdent leur emploi à cause de la propagation de l'automatisation et de l'IA dans tous les secteurs industriels. Une étude bien connue [8] a utilisé un modèle d'apprentissage automatique entraîné sur un échantillon de métiers annotés par des experts du MT pour estimer la probabilité d'automatisation de chaque métier aux États-Unis, en utilisant la taxonomie de la classification type des professions (CTP) comme système de classification.

Ce travail a posé les bases d'une discussion sur le risque de pertes d'emplois dues à l'automatisation. Plusieurs autres études ont suivi celle de Frey et Osborne, notamment les chercheurs M. Arntz, T. Gregory, et U. Zierahn [26] ont étudié les compétences pour estimer le risque de l'automatisation dans les 21 pays de

l'Organisation de coopération et de développement économiques (OCDE). Le Brookfield Institute for Innovation + Entrepreneurship (BII+E), avec le soutien du gouvernement de l'Ontario, au Canada, a étudié l'impact réel de l'IA sur l'emploi (risque de pertes d'emplois dues à l'automatisation et à l'IA) dans le secteur manufacturier et dans le secteur de la finance/assurance, en utilisant les données sur le MT, les publications existantes, des entretiens avec plus de 50 parties prenantes des deux secteurs et la participation de plus de 300 Ontariens par le biais d'entretiens, de consultations publiques et d'une enquête en ligne [27]. L'étude réalisée par T. A. Leopold, V. Ratcheva, et S. Sahiri [28] pour le Forum économique mondial a estimé l'impact de l'automatisation et des progrès technologiques sur le MT, en utilisant des données à la fois sur les métiers et sur les emplois.

En revanche, les entreprises ont besoin d'automatiser les activités de leur département des ressources humaines. En conséquence, un nombre croissant de produits commerciaux dédiés à l'adéquation des compétences ont été développés ces dernières années dans l'UE et hors de l'UE, tels que Burning Glass, Workday, Pluralsight, EmployInsight, Textkernel et Janzz. Il convient de mentionner l'API de recherche d'emploi de Google, un service prépayé annoncé en 2016, qui classifie les postes à pourvoir par le biais du service d'apprentissage automatique de Google sur O*NET, la taxonomie normalisée des professions aux États-Unis.

Matrice projets/caractéristiques: un modèle comparatif pour clarifier quels projets ont abordé une préoccupation/un problème spécifique lié aux Big Data pour l'IMT

Le Tableau 3 résume un certain nombre de projets d'information et de renseignement sur le MT et leurs caractéristiques. Il fournit également une référence pour chaque projet.

Tableau 3 Matrice projets/caractéristiques de l'IMT

Référence	Source(s) de données	Objectif	Pays concernés	Type	Langues traitées
[23]	Groupes administratifs et groupes cibles	Estimer la probabilité d'automatisation	États-Unis	Étude	EN
[24]	Données d'internet (offres d'emploi)	Évaluer l'efficacité de l'IMT sur l'internet pour un suivi en temps réel	Italie, Allemagne, Royaume-Uni, République tchèque, Grèce	Prototype et étude	EN, GR, IT, CZ, DE
[25]	Données d'internet (offres d'emploi)	Développement d'un suivi du MT en temps réel pour l'UE	Ensemble des 28 pays de l'UE	Système et étude de recherche	32 langues
[23]	Enquête et données d'internet (offres d'emploi)	Inclusion de Big Data dans les statistiques officielles	UE	Projet et étude de recherche	S/O

[27]	Données administratives (PIAAC)[9]	Estimer le risque d'automatisation dû à la robotisation et à l'IA	Estimation sur les États-Unis -> s'applique à 21 pays de l'OCDE	Étude	EN
[29]	Données d'internet (offres d'emploi)	Estimer le risque d'automatisation dû à la robotisation et à l'IA	Italie	Étude	EN
[30]	Données administratives (publiées par le gouvernement)	Utilisation des ensembles de données gouvernementales pour mieux comprendre les flux du MT	Royaume-Uni	Projet	EN

Les Big Data pour l'IMT en pratique

Cette partie examine certains projets de travail qui utilisent l'approche ECD décrite précédemment pour mettre en pratique l'IMT à différentes fins. D'une part, nous présentons l'approche WollyBI, qui a été déployée avec succès en Italie et en Espagne (projet Bizkaia).

D'autre part, nous présentons le projet ESSnet Big Data, une initiative de l'UE visant à inclure les Big Data sur le MT dans les statistiques officielles du MT. Ces deux applications mettent en lumière l'importance de l'exploitation des Big Data pour analyser la dynamique et les tendances du MT pour un large éventail de parties prenantes.

D'autres projets et initiatives seront examinés plus en détail dans la partie 3.

Suivi du MT en temps réel Les cas de l'Italie et de l'Espagne

Le projet WollyBI[10] a démarré début 2013 en tant qu'outil SaaS[11] permettant de collecter et de classer les postes à pourvoir publiés sur internet selon les taxonomies normalisées de la classification internationale type des professions/classification des aptitudes, compétences, certifications et professions européennes (CITP/ESCO), et d'extraire les compétences les plus demandées des descriptions de poste. Il a été conçu

9 Programme pour l'évaluation internationale des compétences des adultes (PIAAC). Les données du PIAAC sont une source de données unique qui contient des indicateurs au niveau micro sur les caractéristiques socio-économiques, les qualifications, les informations liées à l'emploi, les tâches et les compétences professionnelles. Il est important de noter que les données sont comparables entre les pays participant au programme. Ainsi, les données permettent également de tempérer l'hypothèse selon laquelle les structures des tâches sont les mêmes dans les différents pays.

10 www.wollybi.com. Droits d'auteur détenus par TabulaeX, une spin-off accréditée de l'Université de Milan-Bicocca, Italie.

11 SaaS signifie logiciel en tant que service (Software as a Service). En tant que service web, accessible à tout moment par toute personne disposant d'un compte valide, il évite de devoir télécharger et installer des outils.

pour fournir cinq points d'entrée distincts pour l'utilisateur en fonction de ses objectifs d'analyse: zone géographique, compétence, entreprise, métier et requêtes gratuites.

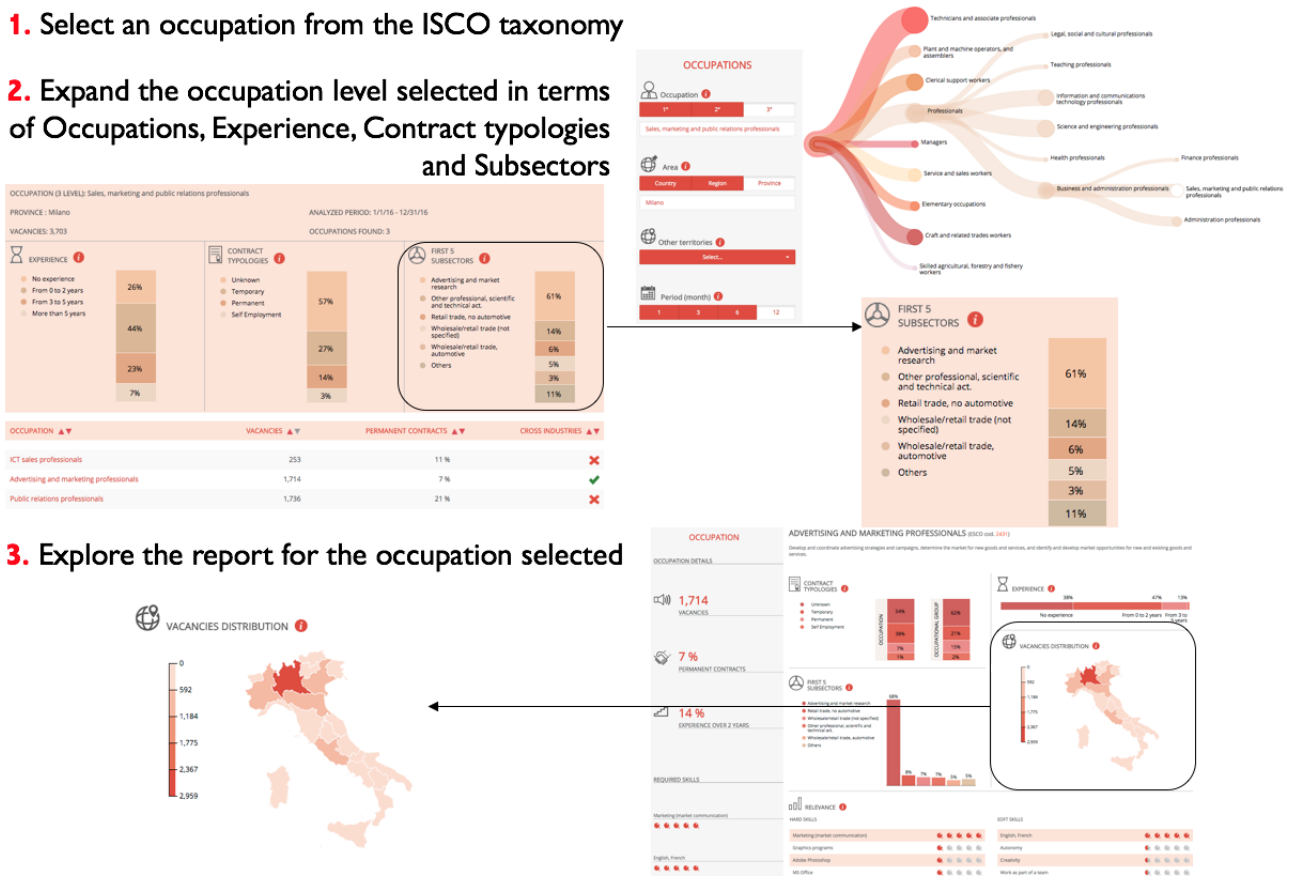
- Zone géographique: trouver les métiers les plus fréquemment recherchés sur le web et les compétences connexes, à un niveau géographique très détaillé;
- Compétence: saisir une sélection de compétences et trouver les métiers les plus fréquemment recherchés qui incluent ces compétences (analyse des écarts de profil);
- Entreprise: obtenir un classement des métiers qui spécifient un secteur industriel particulier dans le poste à pourvoir;
- Métier: naviguer dans les classifications CITP/ESCO et exploiter les détails relatifs à chaque métier;
- Requêtes gratuites (personnalisées): opérations gratuites classiques de zoom avant (drill-down) et de résumé (roll-up) sur les cubes OLAP[12].

La mise en œuvre du projet WollyBI suit de près l'approche ECD, comme l'illustre la référence [15]. Ici, nous nous concentrons sur les chemins de navigation. Ils varient pour chaque partie prenante, mais chaque point d'entrée a été conçu sur la base de la règle des trois clics, de sorte que les résultats sont accessibles à l'utilisateur en un maximum de trois clics, à partir de la page d'accueil.

La Figure 6 montre une séquence de captures d'écran de WollyBI à partir du point d'entrée Métier. Tout d'abord, l'utilisateur doit sélectionner le groupe de métiers du premier au troisième niveau de la CITP et quelques autres paramètres, tels que la zone géographique et l'horizon temporel de l'analyse. Plus la taille des puces est importante, plus le nombre de métiers dans le groupe particulier est important. Un premier rapport indiquant les résultats pour le groupe de métiers choisi est émis. Une fois que l'utilisateur a sélectionné un métier pour l'exploration approfondie, une série d'informations est fournie, notamment l'intitulé du métier, son code ISCO, sa définition selon la taxonomie de la CITP, l'expérience requise, la typologie des contrats, les cinq principaux sous-secteurs et une liste des cinq principales compétences (techniques et non techniques). Une vidéo de démonstration de WollyBI en action sur la dimension Métier est disponible à l'adresse https://youtu.be/zBNsAS5L04g.

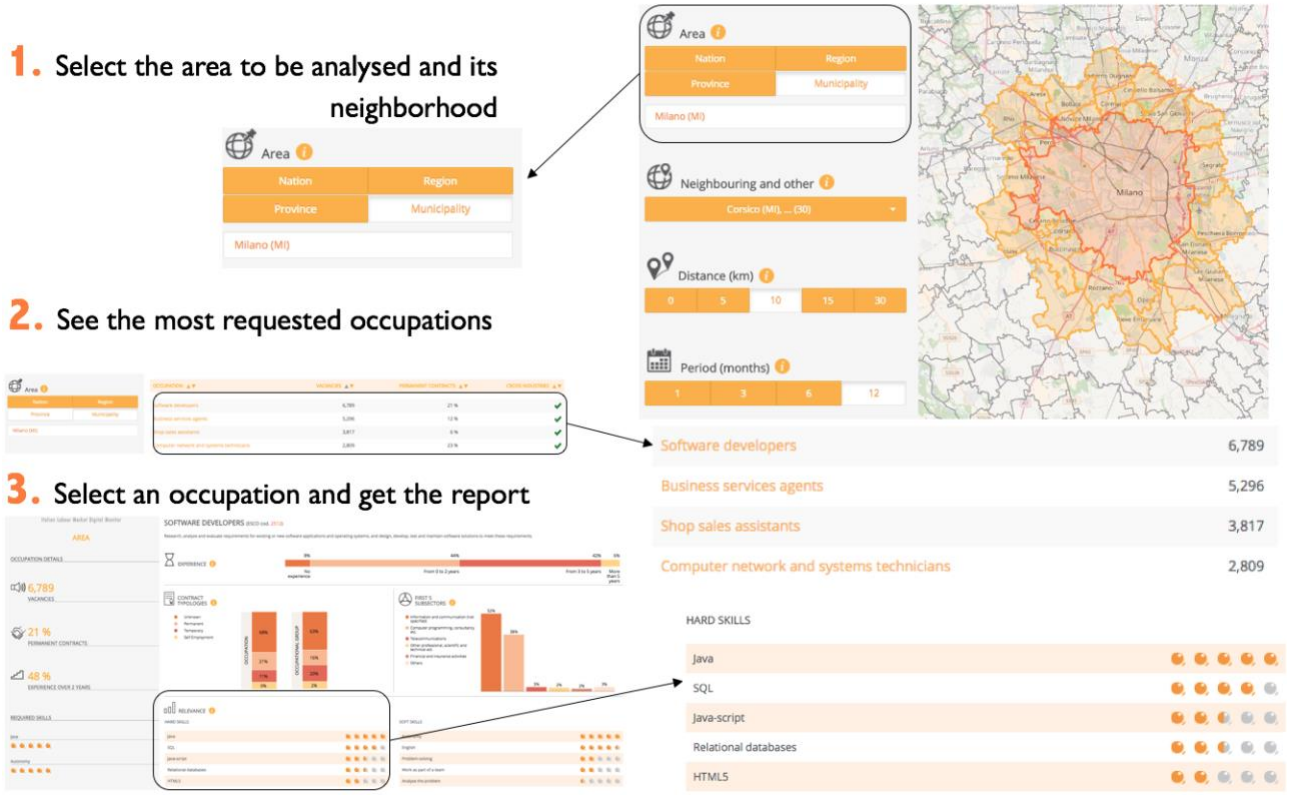
[12] Un cube OLAP est une base de données à plusieurs dimensions, optimisée pour les applications d'entrepôt de données et de traitement analytique en ligne **(OLAP)**.

Figure 6 WollyBI: Séquence de captures d'écran d'une analyse en trois étapes sur la dimension Métier



La Figure 7 montre une séquence de captures d'écran de WollyBI sur la dimension Zone géographique. Tout d'abord, l'utilisateur doit sélectionner la zone géographique qui l'intéresse, l'horizon temporel et un niveau spécifique de la CITP (facultatif). Une liste des métiers de quatrième niveau selon la CITP est fournie, avec le nombre de postes à pourvoir classés selon ce code profession, le pourcentage par rapport au nombre total et une coche indiquant la présence de métiers intersectoriels. Une fois que l'utilisateur a sélectionné un métier, une série d'informations est fournie, notamment l'intitulé du métier, son code ISCO, sa définition selon la taxonomie de la CITP, l'expérience requise, la typologie des contrats, les cinq principaux sous-secteurs et une liste des cinq principales compétences (techniques et non techniques). Il peut s'agir de compétences ne relevant pas de la classification de l'ESCO, telles que spécifiées dans la partie précédente. Une vidéo de démonstration de WollyBI en action sur la dimension Zone géographique est disponible à l'adresse https://youtu.be/Xe_OSOHkx20.

Figure 7 WollyBI: Séquence de captures d'écran d'une analyse en trois étapes sur la dimension Zone géographique



Le projet Bizkaia Talent

WollyBI a été utilisé comme référence – en ce qui concerne les aspects technologiques et méthodologiques – pour le déploiement du suivi du MT au Pays Basque, en Espagne. L'observatoire des talents basques, lancé en 2017, est la première plate-forme mondiale d'accès public pour le suivi du MT hautement qualifié de la région. L'initiative a été élaborée par Bizkaia Talent en collaboration avec TabulaeX, une spin-off de l'Université de Milan-Bicocca, afin de gérer le transfert de connaissances. Elle s'appuie sur un outil qui étudie le MT basque, axé prioritairement sur les professionnels hautement qualifiés, en analysant les Big Data issues de multiples sources en ligne correctement sélectionnées et classées, dans la région du Pays Basque.

L'objectif du projet – soutenu par le ministère de l'économie et du territoire de Biscaye – est de contribuer à la compétitivité de la région de Biscaye et de permettre la collecte en temps réel d'information sur le MT dans la zone géographique de Biscaye, Gipuzkoa et Álava, en utilisant des sources en ligne internationales et locales telles que les universités ou l'agence gouvernementale pour l'emploi LANBIDE. Grâce à l'analyse des Big Data, l'outil crée une base de connaissances sur le MT, la dynamique de l'emploi à un moment donné ou l'évolution des tendances dans le temps, ainsi que sur les exigences techniques et les compétences requises au Pays Basque en ce qui concerne les profils hautement qualifiés. Grâce à des données en ligne mises à jour quotidiennement, il permet aux professionnels hautement qualifiés de suivre les types de profils requis par le MT basque, en fonction de plusieurs types et combinaisons de critères différents, tels que les compétences techniques et transversales requises, le secteur, l'expérience, la zone géographique et le type de contrat. La base de connaissance sur le MT comporte deux points d'entrée distincts:

1. Pour les citoyens: un tableau de bord permettant d'analyser et de parcourir les informations sur le MT sur l'internet en fonction des postes à pourvoir actifs au cours des 12 derniers mois;

2. Pour les analystes: un tableau de bord qui visualise les données des 12 derniers mois, avec une sélection de la période d'intérêt: dernier mois, trois derniers mois, six derniers mois, dernière année. Les données sont collectées quotidiennement et mises à jour mensuellement.

L'outil a été mis à la disposition du public et peut être consulté par tous à l'adresse suivante: https://basquetalentobservatory.bizkaiatalent.eus/visual/public/index#.

Pour plus de détails sur ce projet, voir la partie 3.

Le projet ESSnet Big Data

En se concentrant sur la perspective statistique de l'analyse des Big Data pour l'IMT, Eurostat – l'office des statistiques officielles de l'UE – a lancé un projet nommé **ESSnet Big Data qui vise à intégrer les Big Data relatives à l'information sur le MT dans la production régulière des statistiques officielles**, en utilisant des pilotes pour explorer le potentiel de sources de Big Data sélectionnées et concevoir des applications concrètes [23].

Le projet – lancé fin 2016 – est mené par 22 partenaires de l'UE et suit de près l'approche ECD pour collecter des données à partir de portails web précédemment classés, et nettoyer et transformer les données pour les classer selon les taxonomies normalisées. Il convient de noter ici le rôle important joué par la représentativité des Big Data, car les participants entendent évaluer la capacité des Big Data à être représentatives de l'ensemble de la population (ou d'un échantillon stratifié) en vue de les inclure dans les statistiques officielles de l'UE. Pour autant que nous le sachions, il s'agit de la première initiative publique visant à inclure des Big Data sur le MT (postes à pourvoir par exemple) dans les statistiques officielles, car cela permettra de faire la lumière sur les informations sous-jacentes communiquées par les sources internet, qui doivent être correctement extraites et traitées pour produire une connaissance du MT utile aux décideurs et aux spécialistes du MT.

Q: Qui sont les utilisateurs de ces plates-formes?

R: Les utilisateurs doivent être correctement identifiés au début de la conception du système afin que les connaissances puissent être organisées en fonction des besoins des utilisateurs et de leur capacité à comprendre les données. Par exemple, WollyBI utilise les mêmes connaissances pour différents types de parties prenantes: les agences pour l'emploi, les associations et syndicats d'entreprises, les agences gouvernementales pour l'emploi, les écoles et organismes de formation. Bizkaia Talent est destiné à la fois aux citoyens et aux analystes, tandis que le projet ESSnet s'adresse à l'ensemble des spécialistes et des analystes du MT.

Q: Quelle est la fréquence de mise à jour de ces outils sur le MT?

R: La fréquence de mise à jour doit être décidée en fonction de trois éléments: (i) la mise à jour des sources, (ii) le coût en termes de puissance de calcul nécessaire pour effectuer la mise à jour, et (iii) les besoins des parties prenantes. En général, des mises à jour hebdomadaires sont appropriées à des fins d'analyse.

Q: Dans quelle mesure le processus ECD déployé par ces plates-formes est-il automatisé (IA à 100 %, à 70 %, à 40 %...)?

R: L'intervention humaine diminue au fur et à mesure de l'état d'avancement du processus ECD. Un effort important est nécessaire pour l'identification et le classement des sources, ainsi que pour l'identification des besoins des entreprises et la sélection de bons algorithmes d'IA et des paramètres de réglage correspondants. Une fois ces activités terminées, le système fonctionne de manière autonome, avec des activités de maintenance périodiques. Par exemple, le scraping des données nécessite une plus grande maintenance car le site web peut être modifié de manière imprévisible, alors que le fait d'avoir un accord avec le propriétaire des données permet d'éviter ce problème et de simplifier le scraping. En outre, l'utilisation de l'informatique en nuage (cloud computing) pour réaliser une solution en matière de Big Data réduit considérablement les risques et les coûts liés aux pannes, mais elle pourrait s'avérer coûteuse si la charge de travail (par exemple, le nombre d'utilisateurs ou le traitement par unité de temps) augmente de manière significative. À cet égard, la réalisation d'un SIMT utilisant des Big Data est une approche «human-in-the-loop» (l'humain intervient dans le processus), surtout dans les premières étapes. L'expérience est une compétence cruciale qui peut réduire à la fois les coûts et l'intervention humaine.

PARTIE 2: MESURES SYSTEMATIQUES POUR INTEGRER L'ANALYSE DES BIG DATA DANS L'IMT

Dans cette partie, nous présentons les principaux blocs constitutifs nécessaires au développement d'une architecture «Big Data» (reposant sur des Big Data) pour l'IMT, et examinons certaines conditions préalables importantes: le point de départ, les éléments à prendre en considération, les phases et les principales étapes liées à l'utilisation du processus ECD, évoqué ci-dessus, comme ligne directrice.

Composants nécessaires/recommandés d'une architecture Big Data

Le processus ECD (partie 1, Figure 5) constitue une référence pour la mise en œuvre de toute architecture de gestion des données visant à extraire des connaissances à partir de données, même avant l'ère des Big Data, comme le montre la Figure 8. Chaque étape du processus ECD abordé dans la partie 1 est réalisée pour (i) collecter des données provenant de plusieurs sources de données structurées, et (ii) transformer les données provenant de modèles et formats de données multiples et distincts en un seul modèle de données unifié, en utilisant les techniques et les outils d'extraction, de transformation et de chargement (ETC). Les données sont ensuite généralement stockées dans un entrepôt de données approprié pour des requêtes et des analyses optimisées. Enfin, un tableau de bord permet aux analystes d'effectuer des requêtes et d'utiliser les connaissances restituées sous forme de rapports et de graphiques pour faciliter la prise de décision. Il s'agit d'une architecture d'informatique décisionnelle ou «architecture BI» (Business Intelligence) classique qui fonctionne bien sur des données structurées.

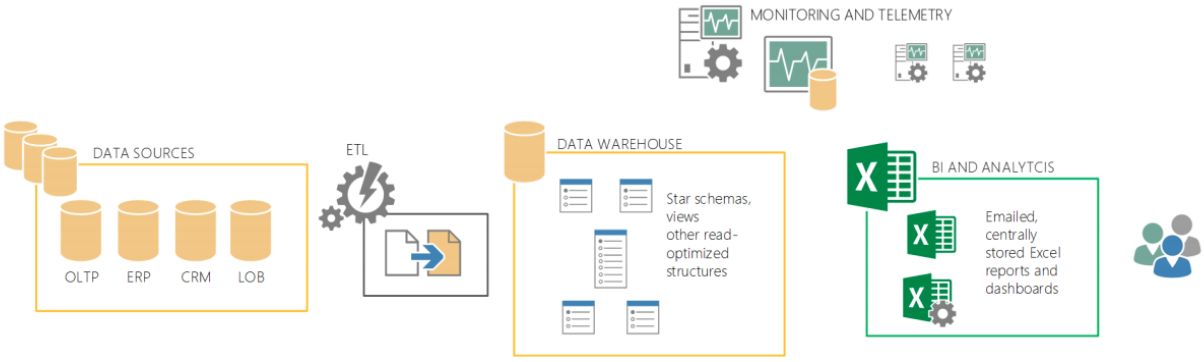
Si ce type d'architecture est très efficace dans de nombreux contextes de la vie réelle, il souffre de limitations qui empêchent leur utilisation dans des scénarios des Big Data, comme le montre le tableau suivant (notamment la colonne «Causé par»).

Tableau 4 Principales limitations de l'architecture Big Data

Problème (le plus important)	Causé par	Éléments conceptuels de l'architecture Big Data
Pas de données sans schéma: seules les sources de données structurées peuvent être manipulées. Cela signifie en gros que seules les données qui obéissent à un modèle de données rigide et bien défini peuvent être manipulées, à l'exclusion de toutes les données «non structurées», telles que le texte libre, les commentaires et le contenu web en général.	Variété	Ingestion de données; modèles NoSQL
Aucune adaptabilité au changement: l'ajout d'une nouvelle source exige que l'ensemble du processus change, ce qui rend difficile de dimensionner l'architecture pour des sources multiples (bien que structurées).	Variété, vélocité	Lac de données
ETC rigides: les procédures qui transforment le contenu des formats sources en formats cibles doivent être rédigées avec précision pour s'adapter à la	Variété	Sans schéma; approche fondée sur les données

structure de données désirée (par exemple, entrepôt de données).		(ascendante plutôt que descendante)
Chronophage: plus le volume de données à traiter est important, plus le processus prend du temps. Les procédures d'ETC sont généralement très chronophages et nécessitent beaucoup de mémoire, car toutes les sources de données doivent être parcourues à tout moment.	Volume, variété, vélocité	Extension horizontale plutôt que verticale

Figure 8 Une architecture BI classique avant l'avènement des Big Data



Le Tableau 4 indique certains des principaux problèmes qui empêchent d'utiliser une architecture BI classique dans le scénario des Big Data, principalement en rapport avec une ou plusieurs dimensions du modèle de Big Data (les cinq V» évoqués dans la partie 1). Ces dernières années, les universitaires et les spécialistes ont déployé des efforts considérables pour identifier des paradigmes, des modèles et des outils adaptés au scénario des Big Data.

Nous présentons ici quelques éléments conceptuels (EC) que le lecteur doit connaître pour mieux comprendre le fonctionnement d'une architecture Big Data; puis, nous examinons comment ces blocs pourraient fonctionner ensemble en vue de réaliser une architecture Big Data pour l'IMT. L'écosystème des Big Data continuant à s'étendre et à évoluer, il existe une multitude de solutions et d'outils.

EC n° 1: Ingestion de données. Ce terme désigne le processus qui consiste à collecter des données provenant de plusieurs sources de manière automatisée. Elle peut être exécutée en **temps réel** (chaque élément de données est collecté alors qu'il est émis par la source) ou via un **traitement par lots** (les éléments de données sont collectés par paquets discrets à intervalles réguliers). En outre, les **données** provenant d'une source unique **peuvent être collectées selon trois méthodes distinctes:** l'**API, le crawling ou le scraping**.

- L'API – comme indiqué dans la partie 1 – désigne une interface de programmation, un composant logiciel mis à disposition de tout programmeur par le propriétaire de la source pour permettre la collecte de données (par exemple, Twitter, Facebook). Le processus de collecte est contrôlé par le propriétaire des données, qui décide également des données qui peuvent être émises, ainsi que de la structure des données.

- Le crawling est une opération logicielle qui consiste à indexer automatiquement tous les contenus d'une page web et à les ajouter à une base de données. Le logiciel («crawler») suit itérativement tous les hyperliens inclus dans la page et indexe également ces données dans la base de données (y

compris les images, les tableaux et les feuilles de style). L'activité de recherche de Google est un exemple classique de crawling.

- Le scraping est une opération logicielle en deux étapes. Dans un premier temps, le logiciel («scraper») demande automatiquement une page web, ensuite, il ne collecte qu'une quantité limitée d'informations de la page et laisse de côté les données restantes. Cela signifie qu'un scraper connaît (partiellement) la structure du site web et peut identifier uniquement le contenu qui l'intéresse pour l'analyse. Par exemple, un crawler peut télécharger tous les produits répertoriés sur un site de commerce électronique, tandis qu'un scraper peut ne collecter que les noms et les prix des produits, en excluant les liens vers les bannières, les commentaires et les métadonnées liés à la mise en page.

EC n° 2: Modèles NoSQL. Ces dernières années, le mouvement NoSQL a fait apparaître de nouveaux paradigmes de modélisation qui diffèrent sensiblement du modèle relationnel classique à la base de toute architecture BI. Quatre paradigmes de stockage de données NoSQL se sont finalement imposés (le modèle clé-valeur, le modèle orienté documents, le modèle orienté colonnes et le modèle orienté graphes). Tous ces nouveaux paradigmes partagent certaines caractéristiques intéressantes par rapport au modèle relationnel classique (voir, par exemple, la référence [31]), notamment un schéma flexible toujours évolutif pour s'adapter aux données, une capacité d'extension horizontale, et un support natif pour le partitionnement. Pour ces raisons, ils sont devenus communs à toute architecture Big Data, car ils permettent de stocker les données sous leur forme originale. Intuitivement, une base de données relationnelle peut se définir comme un ensemble de tables qui peuvent être parcourues à l'aide d'identifiants (ID). Le nombre de colonnes d'une table donnée est fixé et défini a priori. Dans les magasins de données NoSQL, en revanche, le nombre de colonnes peut varier pour chaque ligne de chaque table, ce qui permet de stocker tout type d'élément de donnée, quelle que soit sa structure, car le schéma peut évoluer librement avec les données.

EC n° 3: Lac de données. **Le stockage de données non structurées** (par exemple, le texte libre ou le contenu web en général) **empêche la conception d'un modèle de schéma unifié**. Plus précisément, le schéma des données non structurées peut changer et évoluer dans le temps. Prenons l'exemple d'une tâche de traitement qui consisterait à organiser dans les colonnes d'un tableur commun un million d'offres d'emploi rédigées en texte libre. La difficulté ici réside dans l'identification d'un «modèle» qui correspond à toutes les informations pouvant être extraites des postes à pourvoir. Il est clair que ce modèle peut varier en fonction du schéma des offres. Celles-ci peuvent contenir un nombre variable d'adresses de contact, de lieux ou de compétences, ce qui rend difficile l'identification d'un schéma (ou modèle de données) commun, adapté à toutes les offres, y compris celles qui restent à collecter. Pour remédier à ce problème, **l'une des solutions proposées est l'introduction d'un «lac de données»**. Il s'agit, pour faire simple, d'**un référentiel de données qui laisse toutes les données collectées dans leur format d'origine, en attribuant un identifiant unique à chaque élément de données**. Cela permet de récupérer l'élément lorsqu'il est nécessaire pour effectuer une action donnée (par exemple, analyser les offres de postes à pourvoir de la source X ou du pays Y).

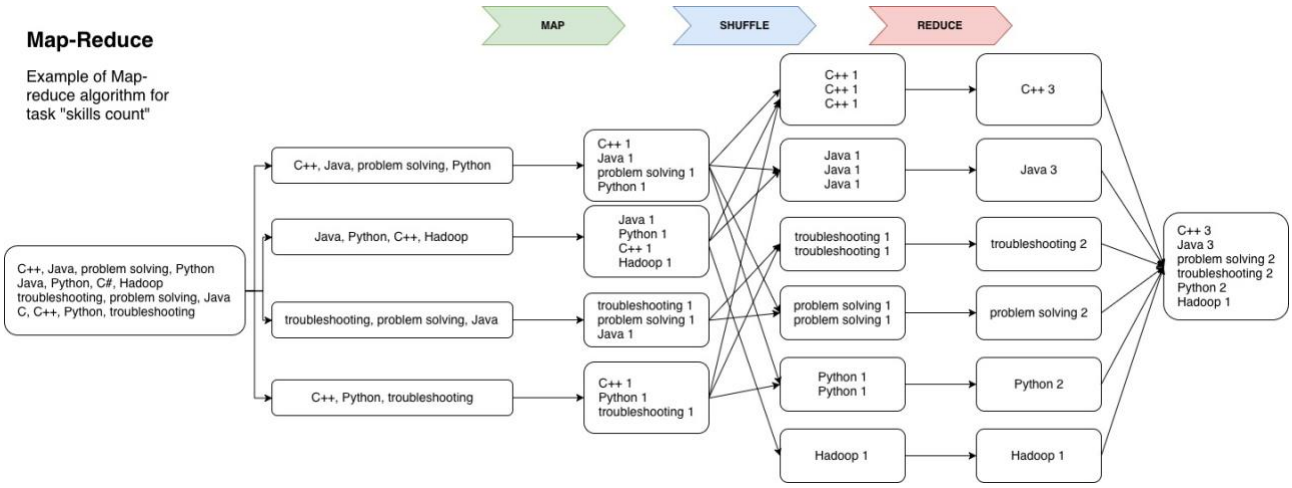
EC n° 4: Extension Globalement, **l'extensibilité signifie que l'architecture peut maintenir le même niveau de performance indépendamment de la charge de travail.** Aucune architecture n'étant pas extensible à l'infini, l'expression «indépendamment de la charge de travail» doit être interprétée comme impliquant des exigences spécifiques d'extensibilité, jugées importantes. Par exemple, une architecture Big Data qui collecte et traite des documents web devrait pouvoir s'ajuster pour garantir une faible latence et un débit élevé. La latence est le temps nécessaire pour effectuer une action (ou produire un résultat) mesuré en unités de temps (minutes pour le traitement en temps réel). Le débit est le nombre d'actions exécutées ou de résultats produits par unité de temps. La latence permet d'assurer que les utilisateurs n'attendront pas indéfiniment la production de résultats, tandis que le débit indique la capacité de l'architecture à traiter les données. Grâce à l'introduction

des magasins de données NoSQL, des systèmes de fichiers distribués[13] et de l'informatique distribuée[14], les architectures Big Data sont extensibles: sur le plan horizontal (et vertical). Pour comprendre comment un système peut réellement s'ajuster, réduire la latence et augmenter le débit, il suffit de comprendre le principe de fonctionnement de MapReduce, moteur central de toute architecture Big Data.

Exemple pratique de MapReduce. Imaginez que des millions de documents arrivent en temps réel de différentes sources internet. Votre architecture a recueilli tous ces documents sous leur forme d'origine en utilisant les magasins de données NoSQL. **L'objectif est de pouvoir traiter tous ces documents en produisant une table de comptage qui indique succinctement tous les mots figurant dans tous les documents ainsi que leurs occurrences**, comme le montre la Figure 9. Le principe de fonctionnement de MapReduce (introduit par [32] alors qu'il travaillait chez Google) vise à répartir la charge de travail entre plusieurs machines, de sorte que chacune ait une tâche très simple à effectuer via la fonction «map». Pour commencer, le fichier d'entrée est partitionné (ligne par ligne) afin que chaque ligne puisse aller à un «mapper» distinct (c'est-à-dire une machine distincte). Chaque machine exécute une fonction très simple («map») qui génère une paire contenant le mot (la clé) et un compteur (la valeur) commençant par 1. À l'étape suivante, chaque paire (ou «tuple») est mélangée (par clé) dans un ordre logique (ordre alphabétique dans ce cas). Intuitivement, le but est que tous les tuples ayant la même clé soient attribués au même «reducer», ce qui permet de gagner du temps. Enfin, chaque *reducer* reçoit, pour chaque clé, toutes les valeurs trouvées dans les tuples. La mission du *reducer* est à présent relativement simple: additionner toutes les valeurs de sa clé, et créer un tuple qui représente le nombre d'occurrences de cette clé, via la fonction «reduce».

Comme le montre la figure ci-dessous, grâce au paradigme MapReduce, quatre machines ont été utilisées pour effectuer une tâche, ce qui a permis de réduire le temps global nécessaire (latence) et d'augmenter la capacité de traitement de millions de documents (débit).

Figure 9 Exemple d'un algorithme MapReduce effectuant une tâche de «comptage des compétences»

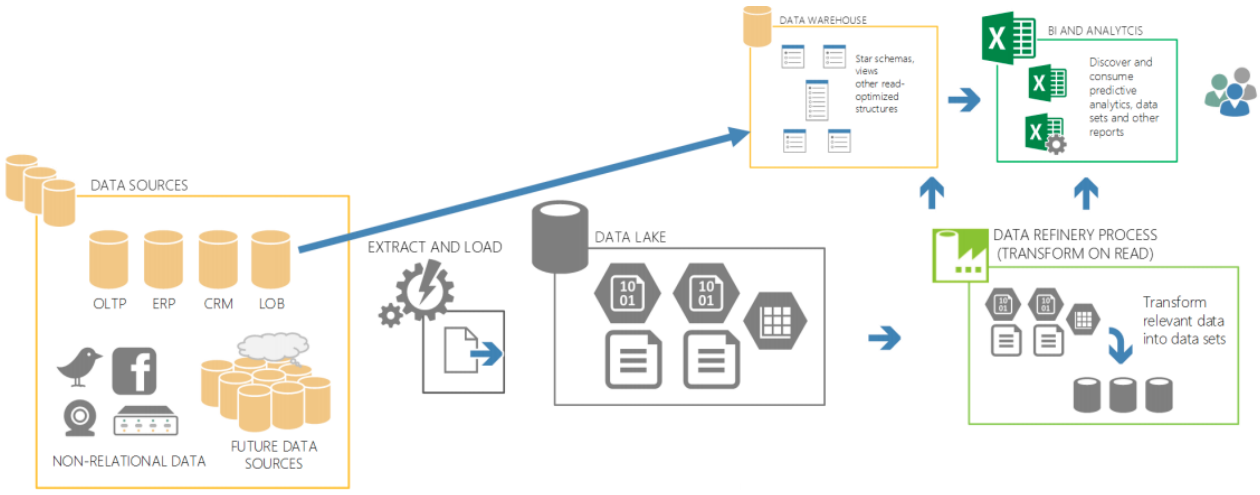


Grâce à cet élément conceptuel introduit par les Big Data, l'approche BI classique présentée à la Figure 8 **peut évoluer vers une approche de Big Data**, comme le montre la Figure 10**Error! Reference source not found.**. En particulier, le flux de traitement de données de l'approche BI classique représenté à la

[13] Les systèmes de fichiers distribués peuvent être considérés comme des structures de stockage sur disque fonctionnant sur plusieurs machines.
[14] L'informatique distribuée désigne la capacité à répartir la charge de travail entre plusieurs machines, réduisant ainsi la charge de travail globale.

Figure 8 existe toujours, les données relationnelles et structurées demeurant présentes dans de nombreux domaines de la vie réelle, y compris les données d'internet. Il existe désormais deux voies distinctes: l'approche classique traite les données structurées, tandis que celle des Big Data, comme nous l'avons vu plus haut, est utilisée pour traiter les données non structurées.

Figure 10 De l'architecture BI à l'architecture Big Data



Architectures, technologies et outils de pointe

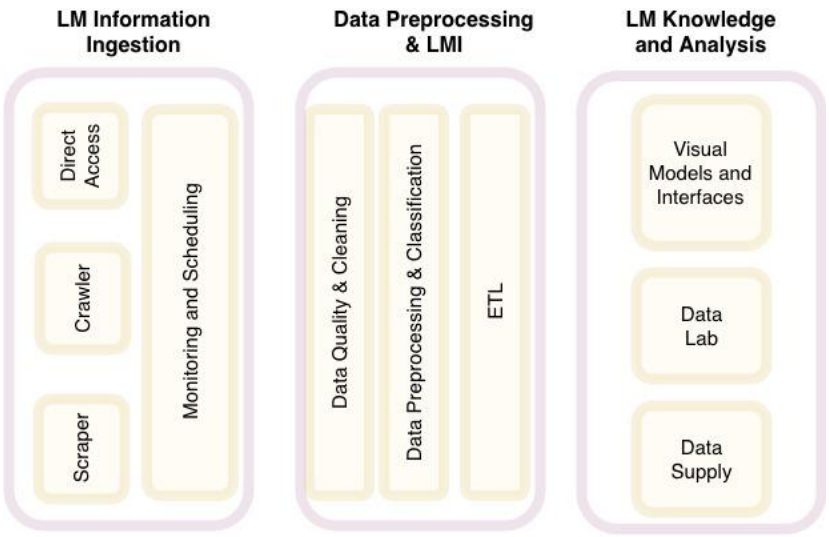
Dans cette partie, nous décrivons les principaux éléments d'une architecture Big Data classique, en présentant des technologies et des outils de pointe. L'écosystème des Big Data s'est développé rapidement ces dernières années[15]; dans ce contexte, nous décrivons ci-dessous les **blocs constitutifs** qui pourraient servir au développement d'une plate-forme de Big Data pour l'IMT.

Architecture Big Data (intuition)

D'un point de vue conceptuel, une architecture Big Data destinée à traiter l'information sur le MT devrait se présenter comme dans la Figure 11. Le but d'une telle architecture est de collecter l'IMT sur l'internet afin d'en tirer des connaissances utiles concernant sa dynamique et ses tendances. Il peut s'agir d'offres d'emploi, de programmes et de données d'enquête sur l'emploi, etc. **Indépendamment de la nature de l'information sur le MT, toute architecture Big Data devrait comporter (au moins) trois macro-étapes**.

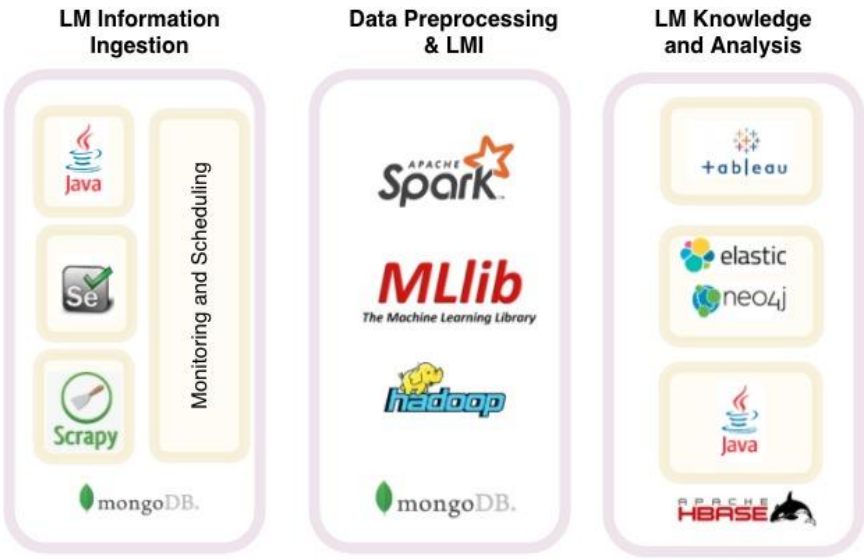
15 Voir https://hadoopecosystemtable.github.io/ – dernière consultation en mars 2019.

Figure 11 Architecture Big Data conceptuelle pour l'IMT en temps réel



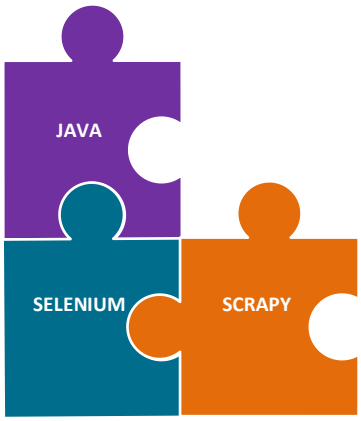
La première étape est la collecte des données, la seconde étape englobe deux tâches d'ECD, à savoir le prétraitement des données et une partie de l'activité de fouille. La dernière étape comprend trois modules principaux pour pouvoir utiliser les connaissances obtenues sur le MT: le module Modèles visuels et Interfaces permet aux utilisateurs finaux de parcourir la base de connaissances de manière interactive en utilisant les paradigmes et les interactions prédéfinies; le Labo de données est un environnement à la disposition des chercheurs pour réfléchir librement sur les données en utilisant des algorithmes d'IA et d'apprentissage automatique pour en tirer des connaissances supplémentaires. Enfin, le module Fourniture de données fournit à des tiers la connaissance du MT qui en résulte. Globalement, cela fonctionnerait comme un service de diffusion de connaissances sur le MT.

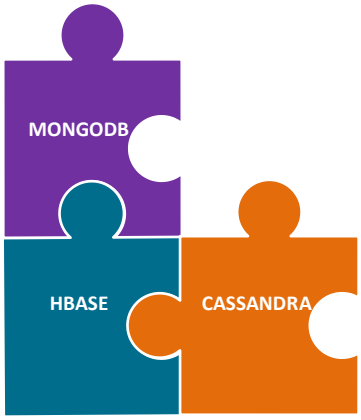
Figure 12 Exemple d'architecture Big Data pour l'IMT en temps réel (écosystème Hadoop)



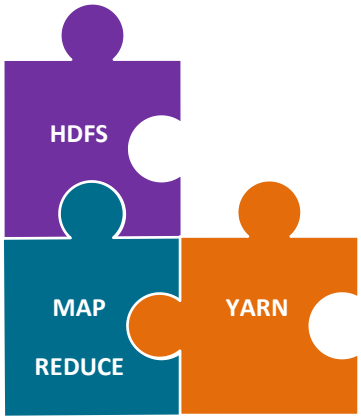
L'architecture suit de près l'approche ECD évoquée dans la partie 1. Ici, nous montrons comment cette architecture conceptuelle peut être déployée en utilisant les technologies et les outils des Big Data. Nous utilisons l'architecture présentée à la Figure 12 à titre d'exemple pour introduire le rôle de chacun des **quatre blocs constitutifs (BC)** de l'architecture Big Data.

BC n° 1: Ingestion d'information sur le MT. Ce module doit être mis en œuvre en tenant compte de la manière dont l'IMT peut être collectée à partir de sources internet, comme nous l'avons vu plus haut. Il existe (au moins) trois catégories différentes de sources d'IMT sur internet : (A) les agences pour l'emploi et les services publics de l'emploi; (B) les journaux, les entreprises et les services d'emploi des universités; et (C) les portails d'emploi. La collecte de données peut varier pour chaque catégorie. Par exemple, (A) pourrait permettre un accès direct, en fournissant des API pour la collecte des données. Dans ce cas, le langage de programmation privilégié devrait être identifié (par exemple Java, Python, Scala) pour se connecter via les API et récupérer les données. Par ailleurs, (B) pourrait ne pas avoir d'API à utiliser. Ces sources modifient fréquemment la structure de leurs pages web, de sorte que la création d'un scraper revient cher et prend du temps. Dans ce cas, un crawler doit être utilisé pour rechercher l'IMT sur le portail. Une solution envisageable serait de recourir à Selenium, outil d'automatisation de navigateur web qui permet de simuler la navigation sur internet et de naviguer dans le contenu web. Enfin, les données de (C) peuvent être collectées en utilisant un scraper; la structure des portails d'emploi ne variant pas souvent, il pourrait être utilisé pour identifier les données collectées à partir de sources internet. Il existe quantité d'outils possibles, tels que Scrapy, plate-forme (*framework*) libre et collaborative basée sur Python pour extraire des données de sites web.





BC n° 2: MongoDB. Il s'agit d'un système de gestion de base de données orienté documents, appartenant à la famille NoSQL. En substance, MongoDB stocke des données structurées sous forme de documents de type JSON[16], plutôt que sous la forme des tables utilisées dans l'approche relationnelle classique. Il permet donc de stocker les données dans leur format d'origine, ce qui constitue un avantage lorsqu'il s'agit de données non structurées dont le schéma peut librement évoluer dans le temps. D'autres systèmes de stockage NoSQL pourraient également être utilisés, tels que HBase et Cassandra, qui sont tous deux des bases de données distribuées orientées colonnes, capables d'effectuer des opérations massives de lecture et d'écriture en temps réel dans de très grandes tables orientées colonnes.



BC n° 3: Prétraitement des données et IMT. Ce module est chargé du prétraitement de l'information sur le MT pour la production de connaissances. Hadoop (2006), plate-forme libre permettant le traitement distribué et le stockage des données pour les applications de Big Data, exécutées dans des systèmes en grappe, marque une étape importante dans le domaine du traitement des Big Data. Le cœur d'Apache Hadoop comprend trois principaux éléments: (i) HDFS (Hadoop Distributed File System), (ii) MapReduce pour l'informatique distribuée, et (iii) la fonction Yet Another Resource Negotiator (YARN) pour la planification des tâches et la gestion des ressources. Ensemble, ces trois éléments permettent de réaliser une architecture distribuée. Si l'on prend comme exemple la classification des offres d'emploi publiées sur internet, le cœur Hadoop permettrait de réaliser:

- le stockage de millions d'offres d'emploi sur plusieurs nœuds, grâce au système de fichiers distribués HDFS. HDFS est très évolutif, fiable et peut utiliser n'importe quelle machine comme matériel de base. Par conséquent, l'information sur le MT peut être stockée sur plusieurs machines;

- les calculs peuvent être distribués sur plusieurs machines grâce à l'algorithme MapReduce décrit précédemment. Dans le contexte de l'IMT, MapReduce peut être considéré comme une fonction de recherche de texte, ou de classement des métiers après leur classification selon une taxonomie normalisée. MapReduce fonctionne très bien sur des ensembles de données très volumineux tels que l'information sur le MT;

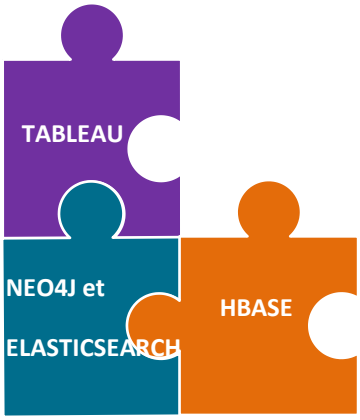
- la planification des flux de travail et la gestion des ressources grâce à YARN. YARN assure la répartition des ressources du système entre les différentes applications fonctionnant dans une grappe Hadoop, et la planification des tâches à exécuter sur les différents nœuds de la grappe.

Apache Spark pourrait être utilisé comme alternative à Hadoop. Le projet Spark[17] est centré sur le traitement des données en parallèle dans une grappe, mais la principale différence est le fonctionnement en mémoire

16 **JSON** (JavaScript Object Notation) est un format léger d'échange de données. Il est facile à lire ou à écrire pour des humains. Il est aisément analysable ou générable par des machines.
17 https://spark.apache.org/ – dernière consultation en mars 2019

vive (RAM). Alors qu'Hadoop lit et écrit des fichiers dans HDFS, Spark traite les données en mémoire vive. À cet effet, Spark a introduit un concept connu sous le nom de «Resilient Distributed Dataset», pour indiquer les ensembles de données qui sont conservés en mémoire. Dans notre exemple d'architecture, nous pouvons utiliser Hadoop et Spark ensemble, Spark pouvant fonctionner en mode autonome, avec une grappe Hadoop comme source de données. L'intégration des deux solutions dans notre architecture pourrait être bénéfique, car elle permettrait de sélectionner le cadre de traitement approprié, en fonction du volume de données collectées. Spark fournit également une puissante bibliothèque d'apprentissage automatique comprenant des algorithmes de pointe, mais toute autre bibliothèque d'apprentissage automatique pourrait être utilisée (voir, par exemple, scikit-learn[18] ou tensorflow[19]).

BC n° 4: Connaissance et analyse du MT. La connaissance du MT générée via l'internet doit être transmise aux utilisateurs finaux en fonction des besoins des parties prenantes. Selon notre expérience, la connaissance du MT sert (au moins) trois parties prenantes/objectifs différents:



1. Les analystes du MT, grâce à des tableaux de bord interactifs qui permettent d'analyser la dynamique et les tendances du MT sur internet selon un schéma prédéfini. Le projet WollyBI par exemple, abordé dans la partie 1, propose quatre différents points d'interaction avec la connaissance du MT, en fonction des besoins des parties prenantes.

2. Les experts du MT, en fournissant un laboratoire de données, environnement de type bac à sable où les chercheurs peuvent appliquer de nouveaux algorithmes et plates-formes pour explorer la synthèse de la connaissance du MT générée via l'internet. Par exemple, la base de données orientée graphes (Neo4j) pourrait être utilisée pour organiser les données comme un réseau social. Si l'on suppose que les nœuds sont des métiers ou des compétences, on pourrait prévoir des indicateurs d'analyse des réseaux sociaux pour trouver des groupes de métiers ou de compétences similaires, ou des groupes de métiers ayant des compétences en commun, ou des analyses de pénuries en vue de recommander les compétences à acquérir pour passer d'un emploi à un autre. De même, les experts pourraient effectuer des recherches textuelles avancées, en mettant en place un moteur de recherche avec ElasticSearch.

3. Les parties prenantes tierces, éventuellement intéressées à utiliser ces connaissances comme un service pour mettre en œuvre leurs propres produits ou services. Cela pourrait être une agence pour l'emploi qui utilise la connaissance du MT pour recommander des postes ou pour soutenir l'enseignement et la formation professionnels (EFP).

D'un point de vue technique, un stockage par famille de colonnes pourrait être utilisé pour une conservation efficace sur les différents nœuds de la grappe, comme dans le cas d'Apache HBase, qui garantit une extensibilité linéaire et des opérations de lecture/écriture en temps réel dans de très grandes tables orientées colonnes. Cela fait partie intégrante de l'écosystème Hadoop, avec la capacité de traitement automatique des tâches MapReduce sur plusieurs grappes.

18 http://scikit-learn.org/stable/ – dernière consultation en mars 2019.
19 https://www.tensorflow.org – dernière consultation en mars 2019.

Le rôle de l'IA pour l'IMT: algorithmes et plates-formes pour l'examen des données du MT

L'intelligence artificielle fait référence à la simulation d'intelligence par des machines. Sa définition a évolué au fil du temps, et la Commission européenne a récemment proposé une explication de qualité: **«L'intelligence artificielle (IA) désigne les systèmes qui font preuve d'un comportement intelligent en analysant leur environnement et en prenant des mesures – avec un certain degré d'autonomie – pour atteindre des objectifs spécifiques.»**[20]. **Cette définition s'applique également à l'IMT, car les algorithmes d'IA** (par exemple pour la classification, la prédiction, la régression et le regroupement) **peuvent être utilisés pour rechercher des tendances intéressantes sous une forme représentative particulière, en fonction de l'objectif de l'analyse**. Plus précisément, comme l'information sur le MT est souvent caractérisée par du texte, les algorithmes d'IA qui travaillent sur des textes pour en extraire des connaissances sont très utiles à cette fin.

Apprentissage supervisé ou non supervisé

Dans le contexte de l'apprentissage automatique, il est utile de faire la distinction entre l'apprentissage supervisé et non supervisé. De manière informelle, l'apprentissage supervisé fait référence aux algorithmes qui permettent d'apprendre à approximer une fonction générique $Y=f(x_1,\ldots,x_n)$ à travers un processus d'entraînement de sorte que lorsqu'un nouvel élément de données d'entrée (x1,…,xn) arrive, le système peut prédire les variables de sortie correspondantes Y. La phase d'entraînement vise donc à faire connaître au système la valeur de Y pour chaque élément de données d'entrée $(x_1,\ldots,x_n)$. L'apprentissage supervisé ne peut donc s'appliquer que lorsqu'une telle fonction cible est disponible dans les données.

L'apprentissage non supervisé concerne les algorithmes dans lesquels il n'y a pas d'information sur la valeur Y correspondante pour chaque élément de données d'entrée $(x_1,\ldots,x_n)$. Par conséquent, l'objectif de l'apprentissage non supervisé est de trouver la structure ou la distribution sous-jacente des données afin d'en apprendre davantage sur les données.

Ce guide ne vise pas à fournir un examen approfondi des algorithmes d'IA (le lecteur est invité à consulter la référence [33] pour une introduction aux modèles d'apprentissage automatique). **Nous présentons donc ici, à l'aide d'un exemple réel en deux étapes, comment des algorithmes et des pipelines d'IA personnalisés peuvent être utilisés pour obtenir des connaissances supplémentaires sur l'IMT**. Dans un premier temps, les algorithmes d'IA (apprentissage supervisé) sont utilisés pour classer les postes à pourvoir publiés sur internet selon les systèmes de classification des professions (comme indiqué à la référence [29]). Dans un deuxième temps, les algorithmes d'IA (modélisation thématique) sont utilisés dans un contexte qui requiert une intervention humaine («human-in-the-loop») pour extraire de nouveaux métiers émergents et déduire le profil du poste en conséquence (voir [34]).

Première étape: Catégorisation de textes via l'apprentissage automatique [supervisé]

Une tâche représentative de IA pour l'IMT s'appuie sur la classification des postes à pourvoir selon une taxonomie normalisée des professions et des compétences. Bien qu'il existe plusieurs taxonomies du MT (par exemple CITP, O*Net, CTP), nous montrons dans cet exemple comment la classification des postes à pourvoir peut être exprimée en termes de classification textuelle. Plus précisément, dans le contexte de l'IMT, il faut généralement utiliser des algorithmes de classification de texte (basés sur l'ontologie,

[20] Commission européenne, L'intelligence artificielle pour l'Europe, COM (2018) 137, disponible à l'adresse https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=COM%3A2018%3A237%3AFIN# – dernière consultation en mars 2019

l'apprentissage automatique, etc.) pour construire une fonction de classification qui affecte un élément de données à l'une des classes prédéfinies. Les éléments sont notamment représentés par des offres d'emploi publiées sur internet, tandis que les classes prédéfinies sont tirées d'une taxonomie [par exemple, une taxonomie d'entreprise ou une taxonomie publique (ESCO ou O*NET) comme indiqué dans l'encadré]. Par conséquent, la tâche de classification des postes à pourvoir peut être formellement définie en termes de catégorisation de texte.

Taxonomie ESCO

L'ESCO est un système de classification multilingue des aptitudes, compétences, certifications et professions européennes, développé par la Commission européenne. La classification de l'ESCO correspond à la Classification internationale type des professions (CITP-08) jusqu'au niveau à quatre chiffres. Elle étend ensuite la CITP en y ajoutant un niveau supplémentaire de professions et de compétences, organisé sous la forme d'un graphe plutôt que d'une arborescence (c'est-à-dire qu'une compétence peut se rapporter à plusieurs professions).

La catégorisation de texte vise à attribuer une valeur booléenne à chaque paire $(d_j,c_i) \in D\times C$, où D est un ensemble de documents et C un ensemble de catégories prédéfinies. Une valeur vraie attribuée à (d_j,c_i) indique que le document d_j doit être classé dans la catégorie c_i, tandis qu'une valeur fausse indique que d_j ne peut pas être classé dans la catégorie c_i. Dans le cas de l'IMT, un ensemble de postes à pourvoir J peut être considéré comme un ensemble de documents, chacun devant être attribué à un seul (et unique) code profession de la taxonomie. La classification d'un poste à pourvoir selon un système de classification implique de lui attribuer un code profession. Cette tâche de catégorisation de texte peut être résolue par l'apprentissage automatique, comme spécifié à la référence [35]. Formellement parlant, si $J = \{J_1,...,J_n\}$ est un ensemble de postes à pourvoir, la classification de J dans la taxonomie consiste en |O| problèmes indépendants de classification de chaque poste à pourvoir sous un code profession donné de la taxonomie o_i pour i = 1,...,|O|. Un classificateur est donc une fonction $\psi : J \times O \rightarrow \{0,1\}$ qui est une approximation d'une fonction cible inconnue $\psi' : J \times O \rightarrow \{0,1\}$. Évidemment, comme ce cas nécessite de traiter un classificateur mono-label $\forall j \in J$, la contrainte suivante doit s'appliquer: $\sum o\in O\ \psi(j,o) = 1$.

Une fois que la tâche de classification des postes à pourvoir publiés sur internet a été modélisée en termes de catégorisation de texte, un algorithme d'apprentissage automatique peut être utilisé pour entraîner le classificateur, dont l'efficacité peut être évaluée à l'aide de plusieurs paramètres. Il va de soi que le classificateur peut s'avérer plus efficace pour certaines classes que pour d'autres.

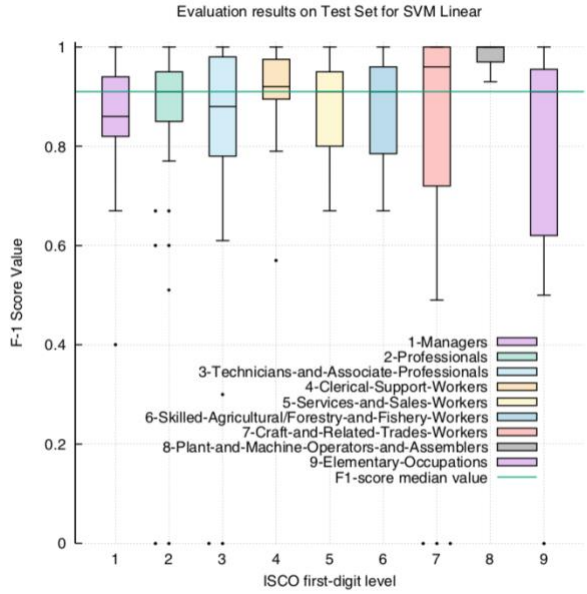
C'est le cas illustré à la Figure 13, où un classificateur d'apprentissage automatique [une machine à vecteur de supports (MVS) en l'occurrence] a été entraîné sur un ensemble de plus de 60 000 offres d'emploi. Le score Formule 1 [21] est indiqué pour chaque premier niveau de la CITP. La boîte à moustaches[22] montre la distribution des valeurs de précision du meilleur algorithme d'apprentissage automatique [c'est-à-dire le classificateur linéaire (MVS)] sur les neuf groupes de la CITP. Ainsi, l'efficacité de chaque algorithme de

[21] Le score Formule 1 (également appelé score F ou mesure F) est l'une des mesures les plus largement utilisées pour évaluer la précision d'un classificateur sur un ensemble de données d'essai.

[22] La boîte à moustaches est une technique statistique bien connue utilisée dans l'analyse exploratoire des données pour identifier visuellement des schémas susceptibles d'être cachés dans un ensemble de données, en mesurant l'évolution des variations entre différents groupes de données. Elle indique la position des quartiles supérieur et inférieur respectivement, ainsi que la valeur médiane, soit la zone comprise entre les quartiles supérieur et inférieur et qui représente 50 % de la distribution. Les lignes verticales (aussi appelées moustaches) s'étendent au-delà des extrêmes de la distribution, indiquant les valeurs minimales ou maximales de l'ensemble de données. Enfin, les points sont utilisés pour représenter les valeurs extrêmes supérieures et inférieures, à savoir les éléments de données qui se situent respectivement à plus (moins) de 3/2 fois le quartile supérieur (inférieur).

classification peut être étudiée sur un groupe spécifique de professions. Bien que le classificateur linéaire ait atteint une précision globale de 0,9, ses performances varient en fonction du premier chiffre de la CITP. **Ces caractéristiques des algorithmes d'apprentissage automatique doivent être considérées avec attention lors de l'évaluation de toute application qui utilise l'apprentissage automatique pour faciliter la prise de décision.**

Figure 13 Rapports détaillés de la précision de l'apprentissage automatique. Extraite de la référence [29]



Deuxième étape: Modélisation thématique pour l'identification de nouvelles professions émergentes [non supervisé]

Nous décrivons ci-dessous comment la modélisation thématique permet d'identifier de nouveaux métiers émergents (potentiels), en exploitant des techniques d'apprentissage non supervisé. **Par «nouveaux métiers émergents (potentiels)», on entend les métiers qui n'ont pas encore été inclus dans un système de classification des professions (c'est-à-dire CITP/ESCO dans le cas présent). L'utilisation d'un nouveau terme lors de la publication d'une offre d'emploi ne signifie évidemment pas qu'il s'agit d'une nouvelle profession, ce nouveau terme devant être confirmé par une tendance croissante qui corrobore la création d'une nouvelle profession (émergente) dans le MT sur l'internet.** À cette fin, l'utilisation de la modélisation thématique est bien adaptée pour identifier les termes qui sont statistiquement significatifs dans les textes.

Plus précisément, supposons une série de documents – composés chacun de 10 ou 15 mots maximum – dont le contenu est un mélange de thèmes (par exemple, thèmes $T_1, T_2, \ldots, Tn$) qui caractérisent le lexique de chaque sous-ensemble documentaire. L'Allocation de Dirichlet latente (LDA [36]) est un modèle probabiliste génératif qui considère chaque document comme un mélange de thèmes latents, où chaque thème est caractérisé par sa propre distribution de mots. Il permet de regrouper les documents par thème, sur la base de la fréquence des mots. Ainsi, chaque thème est composé de mots qui contribuent essentiellement à sa génération. Plus la probabilité qu'un thème contienne un mot donné est élevée, plus la pertinence de ce mot dans le thème correspondant est importante. Le modèle LDA ne fait pas de partitionnement des mots par thème, puisqu'un mot peut relever de plusieurs thèmes, mais avec un degré de pertinence variable.

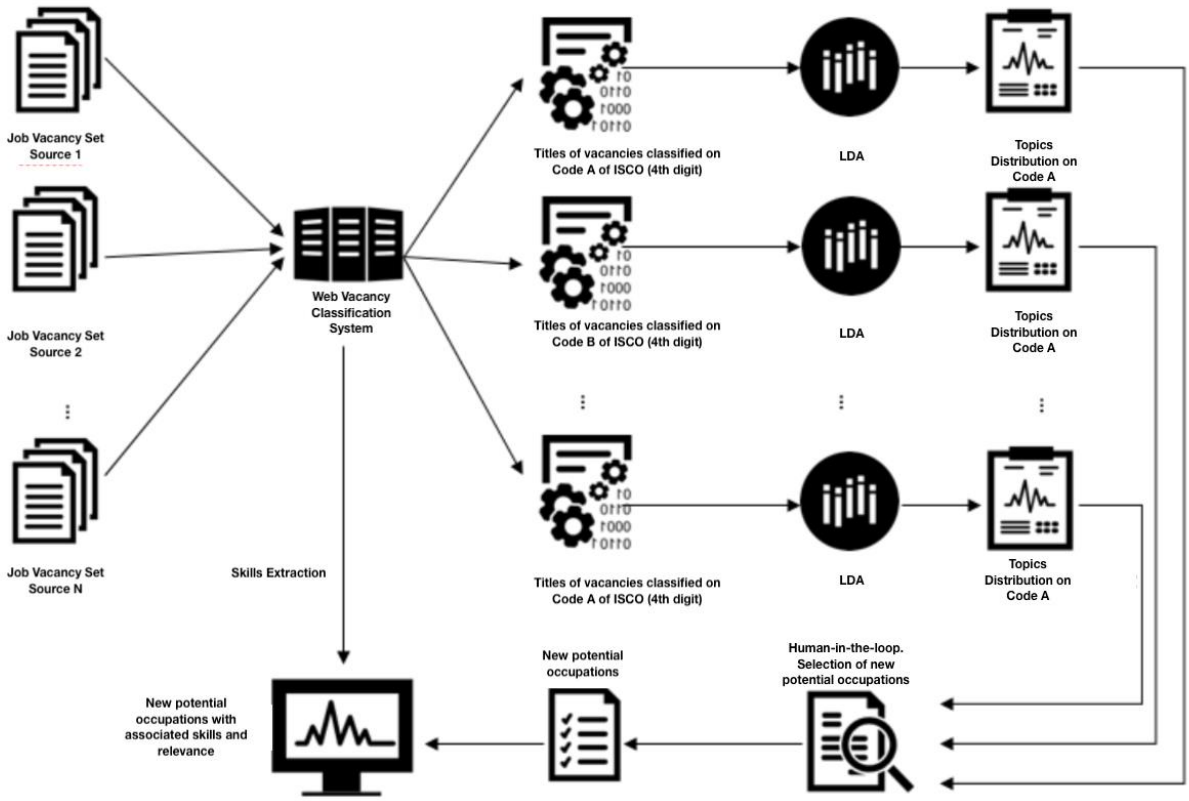
Modélisation thématique (intuition)

Le concept qui sous-tend l'utilisation du modèle LDA pour identifier de nouveaux métiers potentiels consiste à considérer le titre du poste à pourvoir comme un document dont le contenu pourrait idéalement être composé d'un certain nombre (fixe mais inconnu) de thèmes à identifier.

La Figure 14 donne une vue d'ensemble du fonctionnement de ce processus. Une fois que chaque poste à pourvoir a été classé selon la taxonomie normalisée (c'est-à-dire au quatrième chiffre de la CITP en l'occurrence), l'algorithme LDA est appliqué à chaque sous-ensemble de postes à pourvoir, en les regroupant par code CITP. Cette phase de présélection permet au modèle LDA de réduire l'espace des caractéristiques et aussi d'optimiser les performances du LDA. Le processus LDA fournit un certain nombre de thèmes ainsi que la probabilité de distribution des mots qui composent chaque thème[23]. Il délivre une liste de termes principaux pour chaque code CITP, qui doit être analysée et affinée par un spécialiste du MT. **Le LDA étant un algorithme d'apprentissage non supervisé, une intervention humaine pour valider le résultat final est obligatoire afin de garantir la fiabilité des résultats.** Enfin, les mots qui identifient de nouveaux métiers potentiels sont liés aux postes à pourvoir dans lesquels ces mots ont été trouvés, et liés aux compétences incluses. Cela permet d'appréhender de nouveaux métiers émergents et de filtrer les compétences exclusivement demandées par ces derniers.

[23] Le nombre de thèmes à identifier est un paramètre d'entrée de toute approche basée sur le modèle LDA et doit être défini de manière appropriée.

Figure 14 Approche basée sur le modèle LDA en vue de recenser de nouveaux métiers potentiels



À la fin de ce processus, il est possible d'obtenir une fiche de poste, qui identifie:

- **le taux de compétences de la nouvelle profession**, c'est-à-dire le nombre d'occurrences de chaque catégorie ou groupe de compétences, qui peuvent être numériques, non numériques ou non techniques;
- **la répartition des compétences** selon le cadre européen des compétences numériques (e-CF)[24];
- une analyse approfondie permettant de déterminer l'aptitude de l'ESCO correspondante à chaque compétence de l'e-CF telle que trouvée dans les postes à pourvoir.

À titre d'exemple, nous présentons ici le profil du «**spécialiste des Big Data»**. Les données proviennent du projet WollyBI, qui examine les offres d'emploi italiennes sur internet de l'année 2017 seulement, et ont été publiées par l'observatoire italien des compétences numériques[25]. Les compétences numériques ne représentent que 29,5 %, contre 31,5 % pour les compétences non numériques et 39 % pour les compétences

[24] Pour plus d'informations générales concernant l'e-CF, voir le site http://www.ecompetences.eu/. L'e-CF 3.0 peut être consulté à l'adresse http://ecompetences.eu/wp-content/uploads/2014/02/European-e-Competence-Framework-3.0_CEN_CWA_16234-1_2014.pdf.
[25] Accessible au public à l'adresse suivante http://www.assintel.it/assinteldownloads/osservatorio-competenze-digitali-2018-il-volume/ [en italien seulement].

non techniques (Figure 15). La répartition des compétences de l'e-CF est présentée à la Figure 16, qui souligne les compétences exigées en général des «spécialistes des Big Data».

Cet exemple montre comment la connaissance du MT peut être manipulée pour répondre à des questions ad hoc et spécifiques en matière de recherche, telles que l'impact estimé des compétences numériques, non techniques et non numériques dans les professions, l'identification de nouvelles professions émergentes comme celles qui ne sont pas encore codées dans les taxonomies normalisées, et pour comprendre les attentes du MT en se concentrant sur les compétences demandées par le marché, plutôt que sur les compétences figurant dans un profil de poste classique.

Figure 15 Diagramme radar: taux de compétences de la profession de spécialiste des Big Data (tracé vert) par rapport au taux de compétences des autres professions des technologies de l'information et des communications (TIC) (tracé rouge)

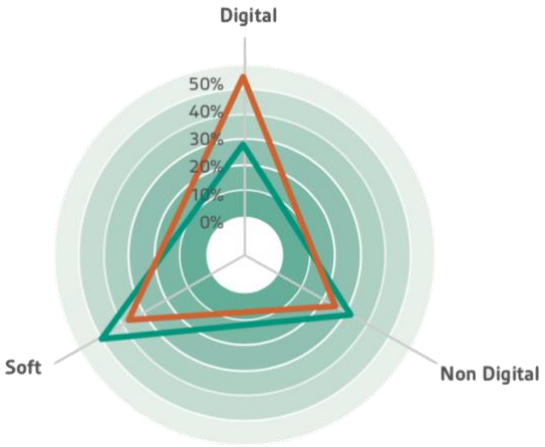
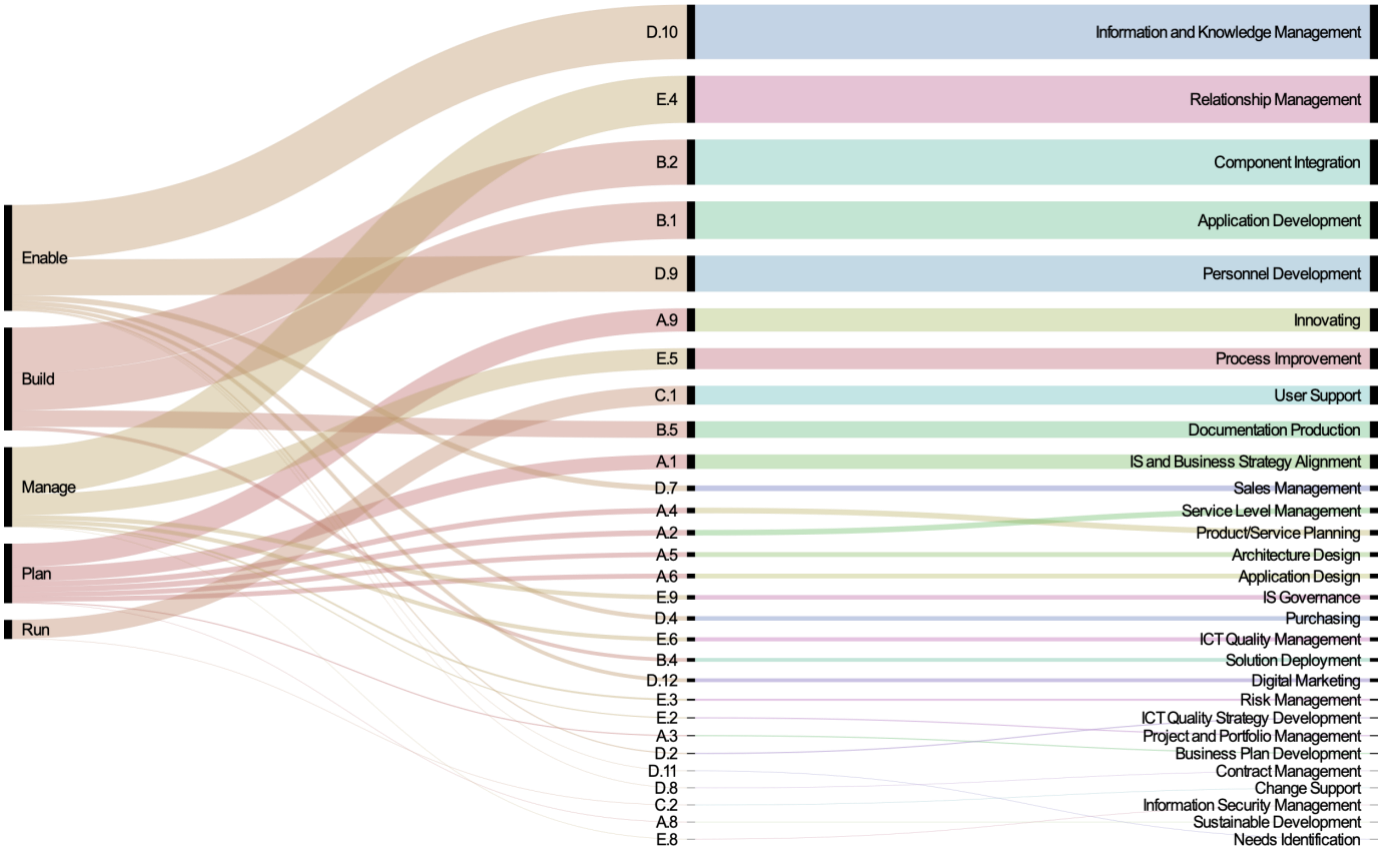


Figure 16 Répartition des compétences de l'e-CF pour la profession émergente de spécialiste des Big Data



Q: Qui écrit les algorithmes d'IA?

R: Jusqu'à présent, les algorithmes d'IA continuent d'être écrits par des hommes. Mais l'idée fondamentale de l'apprentissage automatique est d'avoir des systèmes qui apprennent directement à partir de données. Il n'existe pas de code écrit par l'homme pour guider le processus d'entraînement. Les hommes peuvent seulement essayer de comprendre ce que le système a réellement appris en observant et en interrogeant le système comme une boîte noire.

Q: Qui vérifie la qualité de l'IA pour s'assurer qu'elle fait bien ce qui est attendu d'elle?

R: De nombreux chercheurs travaillent sur ce sujet. L'un des axes actuels concerne l'IA explicable, pour permettre à l'utilisateur de comprendre ce que les machines/algorithmes d'IA apprennent. Il existe à ce jour peu d'outils permettant d'expliquer les critères – et les caractéristiques correspondantes – qui guident un système d'IA dans la prédiction ou la suggestion d'un résultat spécifique.

Q: **Existe-t-il une législation ou des codes d'éthique à l'appui d'une AI de qualité?**

Pas pour le moment, mais en mai 2018, la Commission européenne a publié *L'intelligence artificielle pour l'Europe*, qui examine également certaines lignes directrices en matière d'éthique (voir https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe – dernière consultation en mars 2019).

SECTION 3: EXEMPLES D'UTILISATION DE L'ANALYSE DES BIG DATA POUR LES SIMT: SELECTION DE CAS A UTILISER A TITRE DE REFERENCES PRATIQUES

Dans cette partie, nous faisons le point sur un certain nombre de projets importants, récents, utilisant l'information sur le MT – y compris celle provenant d'applications de Big Data – pour soutenir l'IMT en temps réel, développés en dehors de l'UE (États-Unis et Malawi) et dans l'UE (Royaume-Uni, Pays-Bas, Espagne, et un projet de l'UE impliquant tous les pays de l'UE).

Les éléments présentés ici ont été approuvés et partagés avec les chercheurs principaux des projets respectifs afin d'assurer la description complète de chaque initiative. Chaque projet est donc décrit sur la base des: (i) objectifs ; (ii) données/sources utilisées ; (iii) résultats (ou résultats attendus) ; (iv) réalisations; et (v) questions ouvertes/défis.

CyberSeek.org **[États-Unis d'Amérique]**

CyberSeek.org est une carte interactive de l'offre et de la demande et un portail de carrières pour les emplois dans le domaine de la cybersécurité aux États-Unis. Il s'agit d'un projet conjoint entre Burning Glass, CompTIA et la National Initiative for Cybersecurity Education (NICE). Lancé en 2016, il est financé par le National Institute of Standards and Technology.

Objectifs. CyberSeek vise à fournir des données granulaires et utiles sur le personnel spécialisé dans la cybersécurité aux États-Unis pour aider les éducateurs, les employeurs, les responsables politiques et les demandeurs d'emploi à prendre des décisions mieux éclairées et pour soutenir les efforts visant à combler le déficit de compétences en matière de cybersécurité.

Données et sources utilisées. CyberSeek fournit des données détaillées sur l'offre et la demande d'emplois liés à la cybersécurité dans les secteurs public et privé des différents États et régions métropolitaines des États-Unis. Elles couvrent les paramètres suivants:

- possibilités d'emploi globales;
- nombre total de travailleurs actuellement en activité;
- rapport entre l'offre et la demande;
- quotient de localisation;
- principaux intitulés de postes dans le domaine de la cybersécurité;
- possibilités d'emploi par catégorie du cadre NICE;
- demande et offre de certifications de cybersécurité à la demande;
- nombre total de postes à pourvoir; salaire moyen; compétences, diplômes et expérience requis; et possibilités de transition entre dix emplois de base et cinq emplois d'appoint dans le domaine de la cybersécurité.

CyberSeek tire ces données de trois sources de données principales:

- la base de données de Burning Glass, qui contient des centaines de millions d'offres d'emploi en ligne recueillies depuis 2007. Cette base de données utilise l'analyse avancée du langage naturel pour transformer les informations contenues dans chaque offre de poste à pourvoir en données structurées et utilisables. Elle permet à Burning

Glass de décrire la demande des employeurs pour des fonctions ou compétences spécifiques avec un niveau de détail que les méthodes d'enquête traditionnelles ne peuvent pas offrir;

- les données gouvernementales du Bureau of Labour Statistics concernant la main-d'œuvre actuellement en activité;
- les données sur les titulaires de certificats de cinq organismes de certification: CompTIA, IAPP, ISC^2, ISACA et GIAC.

Résultats (ou résultats attendus[26]). CyberSeek a obtenu trois principaux résultats:

1. la quantification des pénuries de compétences en cybersécurité dans tout le pays;
2. l'identification des possibilités de transition vers et dans la cybersécurité;
3. l'identification des principales compétences et qualifications nécessaires dans le domaine.

Réalisations. Depuis son lancement, CyberSeek a été utilisé par des centaines de milliers d'utilisateurs, notamment des éducateurs, des responsables politiques, des employeurs, des étudiants et des demandeurs d'emploi. Il a été cité par des dizaines de médias comme la source de données incontournable sur le marché de l'emploi dans le domaine de la cybersécurité. Il a également été finaliste du prix du site web à but non lucratif de l'année, décerné par PR Newswire.

Questions ouvertes/défis. Les questions actuelles concernent les défis liés à l'intégration de composantes supplémentaires (par exemple, les étudiants ou les chômeurs) à l'offre de main-d'œuvre existante; la ventilation des données par secteur; l'intégration des données des prestataires de formation; et l'extension de l'outil à de nouvelles zones géographiques.

WheretheWorkIs.org **[UK]**

WheretheWorkIs.org est un modèle de l'offre/la demande d'emplois de niveau intermédiaire, qui identifie les emplois et les régions où il existe des déficits de compétences ou des excédents de travailleurs afin que les prestataires de formation puissent adapter leurs offres en conséquence. Il s'agit d'un projet de Burning Glass Technologies et de l'Institute for Public Policy Research, financé par JPMorgan Chase & Co en 2016 et 2017 dans le cadre du programme New Skills at Work, destiné à identifier des stratégies et des solutions de soutien qui contribuent à améliorer l'infrastructure du MT et à développer globalement la main-d'œuvre qualifiée.

Objectifs. WherheWorkIs vise à fournir un outil unique et gratuit pour permettre aux prestataires de formation, aux chercheurs, aux employeurs et aux décideurs politiques d'examiner les variations de l'offre et de la demande de différents types d'emplois dans plusieurs régions du Royaume-Uni.

Données et sources utilisées. WherheWorkIs utilise deux types distincts de données fournies par Burning Glass:

Données relatives à la demande

Cinquante millions d'offres d'emploi ont été mises en ligne au Royaume-Uni depuis 2012, avec une analyse avancée du langage naturel pour transformer les informations de chaque offre d'emploi en données structurées et utilisables. Cela permet à Burning Glass de décrire la demande des employeurs

26 Plus d'informations à l'adresse suivante https://www.comptia.org/about-us/newsroom/press-releases/2018/06/06/us-cybersecurity-worker-shortage-expanding-new-cyberseek-data-reveals.

pour des fonctions ou des compétences spécifiques avec un niveau de détail que les méthodes d'enquête traditionnelles ne peuvent pas fournir.

La demande de talents de niveau débutant (moins de deux ans d'expérience) est comparée à l'offre disponible de nouveaux diplômés ou de stagiaires. Les données des offres de postes à pourvoir de Burning Glass sont normalisées par rapport aux données sur l'emploi publiées par l'Office for National Statistics (ONS). Elles sont ensuite validées à l'aide de l'enquête annuelle de l'ONS sur le temps de travail et les salaires (ASHE).

Données relatives à l'offre

Burning Glass se fonde sur les effectifs quittant l'enseignement supérieur et la formation continue (les personnes qui terminent les programmes par domaine) pour évaluer «l'offre» de talents débutants. Les données sur l'offre proviennent des organismes suivants:

- Higher Education Statistics Agency (Royaume-Uni);
- Skills Funding Agency (Angleterre);
- Scottish Funding Council;
- Skills Development Scotland;
- Department for Employment and Learning of Northern Ireland;
- StatsWales.

Résultats. Le portail en ligne a révélé des inadéquations considérables entre les besoins en travailleurs moyennement qualifiés des employeurs britanniques pour des postes de premier échelon et les qualifications des nouveaux demandeurs d'emploi – la demande de diplômés de la formation continue et de l'enseignement supérieur est supérieure au nombre de candidats.

Au niveau de la formation, les métiers où l'offre de main-d'œuvre est la plus faible par rapport à la demande sont:

- les services des soins à la personne avec 203 758 postes à pourvoir;
- les vendeurs et les caissiers du commerce de détail avec 89 898 postes à pourvoir;
- les chefs de vente avec 20 139 postes à pourvoir;
- les professions intermédiaires de la santé avec 16 108 postes à pourvoir;
- les métiers du textile et de l'habillement avec 2 905 postes à pourvoir.

Au niveau de l'enseignement supérieur, les métiers où l'offre de main-d'œuvre est la plus faible par rapport à la demande sont:

- les professions de l'enseignement et de l'éducation avec 150 763 postes à pourvoir;
- les métiers des services publics et autres professions connexes avec 71 873 postes à pourvoir;
- les métiers de la garde d'enfants et services connexes avec 29 846 postes à pourvoir;
- les professions de l'aide sociale et du logement avec 20 624 postes à pourvoir;
- les professions intermédiaires de la santé avec 14 390 postes à pourvoir.

L'outil permet aux prestataires d'enseignement et de formation et à d'autres acteurs d'approfondir les données afin de comprendre les opportunités qui existent au niveau local.

Réalisations. Un financement a été accordé pour la mise à jour de l'outil avec des données actualisées sur «l'offre» en 2017. Les réactions des employeurs, des éducateurs et des décideurs politiques ont été positives. Les données utilisées pour cet outil sont à la base des discussions menées entre Burning Glass et un fournisseur d'informations sur les carrières dans les établissements d'enseignement en Angleterre.

Questions ouvertes/défis. Si les données relatives à la demande de Burning Glass sont disponibles en temps réel et suffisamment granulaires pour permettre d'examiner les métiers individuels au niveau local, ce n'est pas le cas des données relatives à l'offre: les données disponibles les plus récentes datent toujours de deux ans lorsqu'elles sont publiées, et l'interrogation au niveau local n'est souvent pas possible en raison d'un nombre insuffisant de données.

Il est souvent difficile de faire correspondre l'offre et la demande. Certains métiers n'ont pas de qualifications ou d'exigences liées à un domaine spécifique (comme la vente, le marketing et les professions connexes). Par conséquent, et bien que de nombreuses personnes possédant un large éventail de qualifications puissent postuler à ces emplois, l'outil les indique souvent comme «sous-dotées», c'est-à-dire avec de faibles «possibilités d'emploi».

Observatoire des talents basques **Espagne:**

Grâce aux Big Data, l'observatoire des talents basques analyse 13 différentes plates-formes d'offres d'emploi publiques et privées, en ligne et hors ligne, centrées sur les postes à pourvoir par des personnes titulaires d'un diplôme universitaire, au Pays basque uniquement, afin de connaître les profils recherchés, avec des compétences techniques et non techniques, sur le marché du travail basque.

Objectifs. Fournir aux professionnels, aux agents publics, au secteur privé et aux universités des informations sur les types de profils, de diplômes et de compétences demandés par le MT, par secteur d'activités, par territoire, etc.

Données et sources utilisées. Le système a analysé 126 000 offres d'emploi au Pays basque au cours des 12 derniers mois, en les filtrant pour garantir une correspondance à 100 % avec deux critères: les personnes titulaires d'un diplôme universitaire, le travail au Pays basque. Les sources internet suivantes ont été utilisées: Adecco, Bizkaia Talent, Université du Pays basque, Indeed, Infoempleo, Jobydoo, Infojobs, LANBIDE (service public de l'emploi du Pays basque), Mondragon People, Université de Mondragon, Monster, Randstad et Studentjob.

Résultats. Une fois au fait des profils professionnels requis, les universités et les agents publics ont pu adapter leur formation à la réalité et aux tendances actuelles, en utilisant les données pour développer un système capable de prévoir les pénuries de talents pour chaque type de profil au Pays Basque, et de mettre en correspondance les offres d'emploi (500 par an) et les professionnels de la plate-forme (11 000), afin de proposer le bon poste au bon candidat et vice versa.

Réalisations. La plate-forme est en accès libre, il suffit de s'inscrire sur www.bizkaiatalent.eus ou www.bebasquetalentnetwork.eus. Les professionnels et les entreprises enregistrés ont également accès à l'application Bizkaia Talent. Pour les nombreux professionnels qui vivent à l'étranger (55 % des 11 000), la connaissance de la demande sur le marché du travail local leur offre de plus grandes possibilités de réinsertion professionnelle dans la région.

Questions ouvertes/défis. Il reste à voir comment utiliser les données et combiner les différentes composantes délais/compétences/métiers pour prévoir la demande de talents dans les années à venir et anticiper les contacts avec les professionnels adéquats.

La taxonomie des compétences fondée sur des données **[Royaume-Uni]**

Nesta[27] a élaboré la première taxonomie des compétences fondée sur des données, mise à la disposition du public au Royaume-Uni. .

La taxonomie a été créée à partir des recherches menées par Nesta en partenariat avec l'Economic Statistics Centre of Excellence (ESCoE[28]). Financé par l'Office of National Statistics (ONS), l'ESCoE a pour objectif de fournir une analyse des questions émergentes et futures concernant les indicateurs de l'économie moderne. Les travaux de recherche de l'ESCoE ont exploré le potentiel d'utilisation des Big Data produites naturellement sous la forme d'offres d'emploi en ligne pour améliorer la compréhension du MT.

Objectifs. La pénurie de compétences est un problème majeur au Royaume-Uni, et susceptible de freiner considérablement la croissance. Selon une étude de l'OCDE[29], le Royaume-Uni pourrait accroître sa productivité de 5 % s'il réduisait l'inadéquation des compétences en s'appuyant sur les meilleures pratiques de l'OCDE.

Les pénuries de compétences, malgré leur importance, ne sont pas actuellement mesurées de manière détaillée et opportune. Les meilleures estimations disponibles proviennent de l'Employer Skills Survey[30]. L'enquête permet de mettre en lumière les différentes causes des pénuries de compétences, mais elle n'est menée qu'une fois tous les deux ans et porte sur des groupes de compétences larges, non détaillés.

L'inadéquation des compétences pourrait s'aggraver à l'avenir, car les compétences nécessaires au travail évoluent en raison de facteurs à court terme comme le Brexit et de tendances à plus long terme comme l'automatisation. Les études de Nesta ont montré qu'un cinquième des travailleurs exercent des métiers qui seront probablement en recul au cours des 10 à 15 prochaines années[31].

La première étape pour mesurer les pénuries consiste à établir une taxonomie des compétences, qui montre les groupes de compétences dont les travailleurs ont besoin actuellement au Royaume-Uni. Cette taxonomie peut ensuite être utilisée comme cadre pour mesurer la demande de compétences parmi les employeurs, l'offre actuelle de ces compétences du côté des travailleurs et l'offre potentielle sur la base des cours proposés par les prestataires d'enseignement ou de formation et les employeurs.

Données et sources utilisées La construction de la taxonomie a commencé avec une liste d'un peu plus de 10 500 compétences individuelles mentionnées dans les descriptifs des 41 millions d'offres d'emploi collectées au Royaume-Uni entre 2012 et 2017 et fournies par Burning Glass Technologies. Ces compétences comprenaient des tâches spécifiques (telles que la souscription d'assurance), des connaissances (biologie), des logiciels (Microsoft Excel) et même des caractéristiques personnelles (disposition positive). L'apprentissage automatique a été utilisé pour hiérarchiser les compétences. Plus la fréquence d'occurrence de deux compétences dans une même annonce était élevée, plus elles avaient de chances de se retrouver dans la même branche de la taxonomie. La taxonomie permet ainsi de déterminer les groupes de compétences nécessaires pour différents emplois.

Résultats

La taxonomie finale a une structure arborescente à trois niveaux. Le premier niveau contient six grands groupes de compétences; ceux-ci sont divisés en 35 groupes, puis à nouveau subdivisés pour donner

27 https://www.nesta.org.uk/ et plus particulièrement https://www.nesta.org.uk/data-visualisation-and-interactive/making-sense-skills/.
28 https://www.escoe.ac.uk/.
29 https://www.oecd.org/eco/growth/Labour-Market-Mismatch-and-Labour-Productivity-Evidence-from-PIAAC-Data.pdf.
30 https://www.gov.uk/government/publications/employer-skills-survey-2017-uk-report.
31 https://www.nesta.org.uk/report/the-future-of-skills-employment-in-2030/.

143 groupes de compétences spécifiques. Chacune des quelque 10 500 compétences se trouve dans l'un de ces 143 groupes. La même méthodologie pourrait être utilisée pour créer d'autres niveaux.

La taxonomie des compétences a été enrichie de manière à fournir des estimations de la demande pour chaque groupe de compétences (sur la base du nombre d'occurrences dans les offres d'emploi), de l'évolution de la demande au cours des dernières années et de la valeur de chaque groupe de compétences (sur la base des salaires indiqués). Les estimations de la demande représentent la moitié des informations sur les pénuries de compétences. Plus important encore, un utilisateur peut effectuer une recherche sur la taxonomie par intitulé de poste et découvrir les compétences nécessaires pour un large éventail d'emplois.

Réalisations. La taxonomie n'a été publiée qu'en août 2018. L'an prochain, toute une série de scénarios utilisant la taxonomie des compétences sera fournie. Il s'agira notamment d'estimer les pénuries de compétences au niveau régional, de détecter automatiquement les ensembles de compétences nouvelles et redondantes, et de déterminer l'offre potentielle de compétences sur la base des cours et formations disponibles. La taxonomie elle-même continuera à évoluer, car le système ajoute un quatrième niveau et tente d'appréhender les relations latérales entre les groupes.

Questions ouvertes/défis. Aucune taxonomie ne saurait être vraiment complète, qu'elle soit élaborée par des experts ou créée à partir d'offres d'emploi, et de plus, il n'existe pas «une bonne façon» de regrouper les compétences. Dans cette étude, la principale limite est que tous les postes à pourvoir ne sont pas publiés en ligne. Par conséquent, la demande de compétences essentiellement utilisées par les travailleurs indépendants ou occasionnels peut être sous-estimée dans la taxonomie. Malgré tout, l'approche fondée sur les données permet d'obtenir la taxonomie des compétences la plus détaillée accessible au public aujourd'hui au Royaume-Uni, et elle est plus facile à actualiser que les taxonomies élaborées par des experts.

Technical, Entrepreneurial and Vocational Education and Training (TEVET) Malawi **[Malawi]**.

L'Organisation des Nations unies pour l'éducation, la science et la culture (UNESCO) et le gouvernement du Malawi ont effectué un examen de la formation technique, entrepreneuriale et professionnelle (TEVET), en 2018. Dans le cadre de l'enquête sur la situation du MT, la demande d'emplois et l'emploi au Malawi ont été analysés en utilisant les Big Data et l'IA afin de démontrer le pouvoir des nouveaux médias et de permettre au Malawi de passer directement à un nouveau système minimisant les frictions sur le MT grâce à des informations en temps quasi réel sur les emplois disponibles et les personnes à la recherche d'un emploi. Le principal site de recherche d'emploi du Malawi, myJobo.com, a vu le nombre de ses utilisateurs doubler tous les dix mois et continue à se développer rapidement.

Données et sources utilisées. Les données sur les postes vacants entre le 1er janvier 2016 et le 30 avril 2018 ont été obtenues sur le principal portail de recherche d'emploi en ligne du Malawi: myJobo.com.

Résultats L'agrégation de Big Data provenant de 360 000 points d'information sur myJobo.com révèle une grande variété de postes, notamment dans l'administration, la comptabilité, la gestion, l'ingénierie, les relations publiques, l'éducation et la santé. Une analyse approfondie des intitulés de poste montre que les emplois les plus demandés en milieu urbain au Malawi entre 2016 et 2018 sont les suivants: comptables, assistants comptables, assistants administratifs, agents financiers, agents techniques, gestionnaires et coordinateurs de projet. Les 100 emplois les plus demandés sont également répertoriés, ce qui donne une image granulaire du marché actuel pour tous les demandeurs d'emploi et

pour les planificateurs gouvernementaux, les prestataires d'enseignement ou de formation, les formateurs et autres acteurs. À titre d'exemple, pour la fonction de comptable, le profil de poste mentionne des connaissances en matière de fiscalité, établissement de rapports financiers, analyse financière, impôt sur les sociétés, audit, budgets et prévisions. Les principales compétences requises pour un comptable sont la comptabilité, les rapports financiers, les budgets, le rapprochement des comptes, Microsoft Excel, le grand livre, les comptes fournisseurs, les contrôles internes, la gestion et l'analyse.

Réalisations. L'analyse des données provenant de myJobo.com fournit des indications importantes sur les tendances des emplois disponibles sur le marché dans une grande partie des zones urbaines du Malawi. Ces tendances apportent de nouvelles informations sur les métiers qui peuvent être quantifiées de manière granulaire pour aider les prestataires de formation à adapter les cours et informer les demandeurs d'emploi sur les types de compétences recherchées par les employeurs. Ainsi, les demandeurs d'emploi et les apprenants peuvent personnaliser leur parcours d'apprentissage afin de répondre à d'éventuelles pénuries de compétences, et des possibilités de perfectionnement peuvent être proposées.

Questions ouvertes/défis. Les possibilités offertes par les nouvelles méthodes utilisant des sites d'emploi constituent un complément utile aux systèmes actuels comme le SIMT, et permettent de combler les lacunes dans la connaissance du MT pour les années comprises entre les enquêtes à grande échelle – comme l'enquête sur la population active du Malawi, menée pour la dernière fois en 2013. Les nouvelles méthodes permettent d'obtenir des informations en temps quasi réel, ce qui aide les demandeurs d'emploi ainsi que les planificateurs de l'éducation et de la formation à se tenir au fait des tendances du marché du travail.

Projets: (A) Professions de transition et (B) indicateurs de tensions **[Pays-Bas]**

Ces projets ont été développés par le service de l'information sur le marché du travail (UWV) (A et B) et Panteia (B) aux Pays-Bas.

(A) Objectifs. Proposer aux demandeurs d'emploi concernés par des professions excédentaires[32] un ensemble de métiers alternatifs et donc plus de chances de trouver un emploi (métiers de transition).

(A) Données et sources utilisées[33]. Le projet s'appuie sur les démarches de mobilité réelles des demandeurs d'emploi par le passé. Il utilise les CV des demandeurs d'emploi sur werk.nl pour la période de 2013 à 2017. Les chômeurs ou les personnes partiellement handicapées qui perçoivent une allocation de l'UWV sont tenus de s'inscrire sur werk.nl et de publier leur CV sur le site. Les personnes bénéficiant de l'aide sociale des communes sont également tenues de s'inscrire sur werk.nl; les demandeurs d'emploi s'inscrivent sur une base volontaire. Le système examine les CV pour déterminer la fréquence des transitions d'une profession à l'autre. En raison d'un éventuel biais de sélection, la base de données ne calcule la fréquence des transitions que pour les professions excédentaires (par exemple, les employés administratifs, les secrétaires, les comptables). Pour ces professions, le projet sélectionne des professions cibles de transition lorsque:

- la profession de transition a de meilleures opportunités d'emploi que la profession excédentaire;

[32] Exemples de professions excédentaires: l'offre de main-d'œuvre dépasse la demande ou bien le nombre de chômeurs de courte durée est plus de 1,5 fois supérieur au nombre de postes à pourvoir.
[33] Les données peuvent être (semi-)structurées/non structurées ; les sources peuvent être administratives, statistiques, provenir d'enquêtes, de sites web.

- le niveau de la profession de transition est (quasiment) similaire à celui de la profession excédentaire.

(A) Résultats. Pour les professions excédentaires, le système publie périodiquement plusieurs professions de transition. Les informations sont utilisées par les conseillers pour l'emploi, ainsi que directement par les demandeurs d'emploi. Elles sont également utilisées dans les présentations, les ateliers et les webinaires.

(B) Objectifs. Fournir un solide indicateur de tensions sur le MT par région et par groupe de professions.

(B) Données et sources utilisées. Le projet utilise une combinaison de Big Data, de données d'enquête et de données administratives:

Les Big Data exploitent les offres d'emploi en ligne telles que celles de la base de données Jobfeed de Textkernel, après avoir éliminé les doublons et effectué des contrôles de fiabilité.

Les données d'enquête comprennent les résultats de l'enquête nationale sur les postes à pourvoir du bureau central des statistiques afin d'assurer la cohérence et de tenir compte des biais de sélection par secteur et par niveau d'éducation.

Les données administratives couvrent les personnes recevant une allocation de chômage de l'UWV depuis moins de six mois.

(B) Résultats. Tous les trois mois, le système publie l'indicateur de tensions pour 35 régions dans 114 groupes professionnels et indique les types suivants: pénurie très importante, pénurie, dans la moyenne, excédent, excédent important.

Le système est utilisé pour les décideurs politiques, la presse, les présentations, les ateliers, les contributions à plusieurs publications et d'autres objectifs de recherche.

Informations en temps réel sur le marché du travail concernant les compétences requises **[tous les États membres]**

Après une étude pilote concluante, le Cedefop[34] a commencé en 2017 à développer un système de collecte de données sur les offres d'emploi en ligne dans tous les États membres de l'UE, avec le lancement de l'appel d'offres intitulé «*Real-time labour market information on skill requirements: setting up the EU system for online vacancy analysis*». Le système doit être pleinement développé et opérationnel d'ici 2020.

Le Cedefop coopère avec les organismes suivants pour développer le projet:

- CRISP – Université de Milan-Bicocca (Leader): Le CRISP (centre de recherche interuniversitaire sur les services publics) est un centre de recherche basé à l'université de Milan-Bicocca (ITALIE);
- TabulaeX s.r.l.: une spin-off de l'université de Milan-Bicocca (ITALIE);
- IWAK (Institut pour l'économie, la culture et le travail): un institut de recherche appliquée associé à l'université Goethe de Francfort-sur-le-Main (ALLEMAGNE).

[34] Le Cedefop est l'une des agences décentralisées de l'UE. Créé en 1975, le Cedefop apporte son soutien à la Commission européenne, aux États membres et aux partenaires sociaux dans l'élaboration de la politique européenne en matière d'enseignement et de formation professionnels (EFP) et contribue à sa mise en œuvre.

Le projet impliquant le développement d'un système de classification multilingue, le consortium collabore également avec des experts nationaux et internationaux représentant les 28 États membres de l'UE.

Objectifs. Ce projet a pour principal objectif de développer un système à grande échelle permettant au Cedefop de procéder à l'analyse des offres d'emploi en ligne et des besoins en compétences émergentes dans les 28 États membres de l'UE. Le système de collecte de données en temps réel permettra de recueillir les informations de base pertinentes sur les emplois, les entreprises et les types d'employés recherchés (compétences, qualifications et autres caractéristiques) afin de faciliter les futures recherches et analyses de la demande de compétences.

Il sera créé à partir d'outils d'ingestion de données en ligne qui permettent de consulter systématiquement les sites et les éléments impliqués dans l'analyse programmatique des pages web afin de recueillir des formes spécifiques d'informations. Le résultat final sera une base de connaissances qui fournira des informations sur la demande de main-d'œuvre, en soulignant notamment les compétences requises. Cette base de connaissances pourrait être utilisée pour effectuer de nombreux types d'analyses en vue d'aider les parties prenantes et les décideurs politiques.

Le processus comprend le nettoyage des données des offres de postes à pourvoir publiées sur internet et la classification des variables extraites. Les résultats de l'analyse statistique seront rendus accessibles par des outils de visualisation.

Données et sources utilisées[35]**.** La première phase du projet, «Landscaping activity» (activité d'inventaire), a examiné les sources de données disponibles dans l'UE. L'objectif était de comprendre l'utilisation des offres d'emploi en ligne par les employeurs et les demandeurs d'emploi, et d'évaluer la représentativité des données pour une interprétation adéquate des résultats. Cette phase a également permis d'établir une liste de portails web appropriés pour les activités d'ingestion de données.

La liste des sources pertinentes pour l'ensemble des 28 États membres de l'UE indiquait 530 sources, réparties dans les catégories suivantes:

- moteurs de recherche d'emploi:
- agences pour l'emploi;
- sites web sur l'emploi;
- portails d'annonces classées;
- entreprises;
- sites web de services publics de l'emploi;
- journaux en ligne;
- enseignement;
- organismes pour l'emploi.

35 Les données peuvent être (semi-)structurées/non structurées ; les sources peuvent être administratives, statistiques, provenir d'enquêtes ou de sites web.

Les méthodes d'ingestion des données varient selon les sites web: scraping (24 %), crawling (18 %) et accès à une API (57 %), réalisés avec les sites web les plus importants dans le cadre d'accords de licence relatifs à l'utilisation des données.

Après les six premiers mois d'ingestion, le projet a estimé à environ 60 millions le nombre de postes individuels à pourvoir par an pour l'ensemble des pays de l'UE.

L'ingestion d'offres d'emploi en ligne visait à extraire les variables suivantes, classées comme indiqué:

- profession--> ESCO v.1 jusqu'au niveau 4;
- niveau d'enseignement--> Classification internationale type de l'éducation (CITE) niveau 1;
- territoire --> NUTS[36] jusqu'au niveau 3;
- temps de travail --> taxonomie personnalisée («temps partiel» et «temps plein»);
- type de contrat --> taxonomie personnalisée («temporaire», «permanent», «indépendant», «stage»);
- industrie --> niveau NACE[37] jusqu'au niveau 2;
- salaire --> taxonomie personnalisée basée sur des classes numériques;
- compétence --> ESCO v.1.

Résultats Le projet débouche sur trois résultats:

1. réalisation de l'exercice d'inventaire (*landscaping*) dans chaque État membre;
2. réalisation d'un système complet de collecte et d'analyse des offres d'emploi en ligne dans les 28 États membres de l'UE;
3. contribution à la production d'un rapport analytique basé sur les résultats et à sa diffusion ultérieure.

Réalisations. Le projet se poursuit, la publication anticipée des résultats couvrant sept pays est prévue au premier trimestre 2019.

Questions ouvertes/défis. L'appui à de nouveaux modèles d'analyse fondés sur les Big Data et les offres d'emploi en ligne se heurte à des difficultés; il convient d'inciter les décideurs politiques et les parties prenantes à adopter une approche de prise de décision fondée sur les données. Il convient de développer des tableaux de bord de visualisation personnalisés par type d'utilisateurs (décideurs politiques, statisticiens, etc.), et une architecture extensible s'impose pour prendre en charge des volumes de données croissants. Il reste des défis à relever en ce qui concerne la ventilation des données sur les compétences en fonction des classifications spécifiques (compétences en TIC, compétences non techniques, etc.), et des efforts supplémentaires à consentir pour soutenir de

36 Nomenclature des unités territoriales statistiques, voir https://ec.europa.eu/eurostat/fr/web/nuts/background

37 Nomenclature statistique des activités économiques dans la Communauté européenne – voir: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Statistical_classification_of_economic_activities_in_the_European_Community_(NACE)/fr

nouvelles études et recherches sur les tendances de l'emploi axées sur les données des offres d'emploi en ligne.

SECTION 4: CONCLUSIONS ET RECOMMANDATIONS.

Résumé des recommandations et des mesures à l'attention de l'ETF et de ses pays partenaires

Le développement éventuel d'un système d'analyse de la demande de main-d'œuvre sur internet implique l'examen d'un certain nombre de macro-sujets:

INFORMATIONS GÉNÉRALES Les objectifs et les caractéristiques du système peuvent être décrits de la manière suivante:

1. TERRITOIRE D'INTÉRÊT: Quel MT souhaitons-nous étudier? Un seul pays, un groupe de pays ou un continent? Nous intéressons-nous à l'ensemble du MT ou à un seul segment ou secteur? Le choix aura une incidence sur les taxonomies utilisées dans le projet, car il est nécessaire de trouver un terrain commun pour comparer les résultats.
2. VARIABLES D'INTÉRÊT: Quel type de variables voulons-nous étudier? Si nous souhaitons créer une base de connaissances axée sur la demande de main-d'œuvre, c'est-à-dire utiliser les offres d'emploi en ligne comme données, nous devons décider: s'il est suffisant de se concentrer sur les professions; si nous voulons élargir l'analyse pour y inclure les compétences (en reliant les compétences requises et les professions en question) ou si nous voulons exploiter les données des offres d'emploi en ligne pour en extraire le pouvoir informatif et analyser les niveaux d'éducation, les niveaux territoriaux, le temps de travail, le type de contrat, le secteur d'activité et le salaire. Cette décision aura bien sûr une incidence sur le budget en fonction de l'espace de stockage nécessaire et du temps consacré à l'approche méthodologique.

PROCESSUS DE SÉLECTION DES SOURCES. Une fois la décision prise d'utiliser les offres d'emploi publiées en ligne, la principale question est de savoir sur quels sites web il faut les collecter. Cette phase pourrait comprendre les volets suivants:

1. ENQUÊTE PRÉALABLE GLOBALE SUR LES SITES WEB. Celle-ci couvre de nombreuses questions clés, telles que l'utilisation des offres d'emploi en ligne par les employeurs et les demandeurs d'emploi (un point important pour évaluer la représentativité et la crédibilité des données) et le scénario local général. Ces questions doivent être traitées par des experts qui connaissent le contexte et les pratiques les plus courantes.
2. MÉTHODOLOGIE DE SÉLECTION DES SOURCES. Comme nous l'avons déjà mentionné, il existe de nombreuses sources disponibles. Après l'enquête préalable, nous devons décider si nous voulons prendre en considération chaque type de source ou s'il existe des critères d'exclusion.
3. CLASSEMENT et SÉLECTION DES SOURCES. Le classement peut être établi en utilisant un modèle statistique très sophistiqué ou bien un modèle simple qui prend en considération les variables nécessaires pour pouvoir exclure automatiquement les sites web ne possédant pas ces attributs. Par exemple, compte tenu de l'objectif d'analyse en temps réel, un attribut fondamental de la source pourrait être la mise à jour: les sites web non actualisés depuis longtemps (plus de six mois, par exemple) pourraient être exclus. Le résultat du classement serait une liste des sources les plus importantes: à ce stade, il faudrait décider de les prendre toutes en considération ou de faire une sélection en fonction de la capacité de performance.

QUESTIONS ÉTHIQUES ET JURIDIQUES. La législation relative à l'utilisation des données et, plus précisément, des offres d'emploi en ligne n'est pas toujours très claire. Il conviendrait d'effectuer des

recherches pour identifier précisément les domaines d'action et les activités qui sont expressément interdites.

L'APPROCHE DE BASE DE DONNÉES ET LE MODÈLE DE DONNÉES. Avant de commencer l'ingestion d'offres d'emploi en ligne à partir des sites web sélectionnés, nous devons décider comment traiter toutes les données collectées au cours du projet et établir une solide méthodologie sur le plan technique

1. APPROCHE DE BASE DE DONNÉES: Dans le domaine de la technologie des bases de données, il existe deux grands types de bases de données: SQL et NoSQL (autrement dit, les bases de données relationnelles et les bases de données non relationnelles). La différence réside dans leur structure, le type d'informations stockées et les modalités de stockage. Les bases de données relationnelles sont structurées, à l'instar des annuaires téléphoniques qui contiennent des numéros de téléphone et des adresses. Les bases de données non relationnelles sont «orientées documents» et distribuées, comme des dossiers contenant toutes les données d'une personne: adresse, numéro de téléphone, «likes» de Facebook, préférences d'achat en ligne, ... Nous devons choisir le type de base de données à utiliser pour le stockage des offres d'emploi en ligne.
2. MODÈLE DE DONNÉES: le modèle de données dépend du contenu de l'offre d'emploi et des dimensions à prendre en compte. Le modèle de données source comprend les structures alimentées par les processus d'ingestion et les résultats de la phase de prétraitement. Le modèle de données d'étape comprend les structures utilisées pour stocker les résultats de la phase de catégorisation. Les métadonnées comprennent toutes les structures utilisées pour stocker les ontologies, les taxonomies et les collections de services. Le modèle de données de présentation comprend toutes les structures mises à la disposition des utilisateurs à des fins d'analyse.

LE PROCESSUS D'INGESTION DE DONNÉES. Après avoir identifié les sources d'offres d'emploi en ligne, nous devons décider comment collecter les données. Les techniques d'ingestion les plus courantes sont le scraping, le crawling et l'accès direct aux données par le fournisseur, qui peut proposer une connexion à une API ou fournir directement les données dans un format convenu. L'accès à une API faciliterait grandement la mise en place du système, en accélérant et améliorant l'efficacité de l'ensemble du processus technique. Dans tous les cas, il serait souhaitable de contacter les webmasters des sources sélectionnées pour les informer du projet et déterminer l'approche la plus appropriée pour le téléchargement des données.

PRÉTRAITEMENT. Il s'agit d'une étape d'une importance cruciale, intervenant après la collecte des offres d'emploi en ligne. Au cours de cette phase, les données sont nettoyées pour éliminer le bruit et préparées pour l'étape d'extraction d'information. Il y a également plusieurs étapes d'évaluation de la qualité des données. L'une des questions qui se posent pendant cette phase est celle de la déduplication: dans quelles circonstances une offre d'emploi en ligne doit-elle être traitée comme le doublon d'une autre offre d'emploi en ligne?

EXTRACTION ET CLASSIFICATION DE L'INFORMATION

1. TAXONOMIES: des taxonomies doivent être choisies pour classer les variables. Plusieurs options sont possibles: pour la majorité des variables, il existe des classifications normalisées adoptées dans les pays et développées au fil des ans (par exemple, ESCO pour les professions et les compétences, NUTS pour le niveau territorial, NACE pour les industries). Si le MT en question concerne plusieurs pays, il est important de garantir des taxonomies communes pour toutes les variables.
2. TECHNIQUES D'EXTRACTION DE L'INFORMATION ET AFFECTATION DE TAXONOMIES: pour décider des techniques d'extraction d'information à adopter, il y a lieu de se pencher sur l'état

de l'art afin de choisir entre les systèmes basés sur l'ontologie ou les systèmes d'apprentissage automatique (supervisé ou non).

ETC ET PRÉSENTATION La dernière étape consiste à décider de la manière dont l'information sera affichée et visualisée. Il conviendra ensuite de concevoir le modèle de présentation, les chemins de navigation et les tableaux de bord, et d'évaluer les types d'utilisation.

Idées de projets pilotes

Dans cette partie, nous récapitulons les principaux résultats du groupe de travail sur l'utilisation des Big Data pour l'IMT lors de l'atelier «Skills needs uncovered: dealing with uncertainty», dans le cadre de la conférence «Compétences pour l'avenir: gérer la transition», organisée par l'ETF à Turin les 21 et 22 novembre 2018.

Cet atelier ainsi que les activités menées par les groupes de travail ont permis de discuter et d'échanger des idées, des expériences, des techniques, des défis et problèmes de la vie réelle, et de favoriser l'enrichissement mutuel entre les chercheurs et les spécialistes travaillant sur le MT et les Big Data en vue d'identifier de nouvelles orientations, actions et idées de projets en la matière. Plus précisément, les discussions entre les participants ont porté principalement sur quatre sujets distincts: (i) les avantages pour les citoyens en termes de parcours professionnels; (ii) le rôle des Big Data dans le contexte de l'IMT; (iii) le renforcement de l'utilisation des Big Data dans les pays en développement et en transition; et (iv) des idées de projets pilotes. Nous présentons ci-dessous les principaux aspects liés à chaque sujet qui ressortent des activités des groupes de travail et ont été pris en compte pour élaborer des idées de projets pilotes.

Les avantages pour les citoyens en termes de parcours professionnels

L'utilisation des Big Data à l'appui des parcours professionnels est l'une des applications les plus intéressantes (et les plus perturbantes) de l'IMT. Elle offrirait certains avantages aux citoyens, tels que:

- la possibilité d'analyser leurs lacunes, c'est-à-dire les écarts entre les compétences qu'ils possèdent et les compétences demandées par le MT;
- la possibilité de classer les professions selon une taxonomie normalisée (ESCO, O*NET ou CTP) qui permet de comparer le même emploi dans différents pays, favorisant ainsi la mobilité entre les États membres de l'UE;
- la création d'un outil de service dédié aux carrières;
- la sensibilisation aux changements rapides que connaît le MT et à l'importance de l'apprentissage tout au long de la vie;
- la possibilité de relier leur propre parcours d'apprentissage à un parcours professionnel et aux compétences émergentes.

Le rôle des Big Data dans le contexte de l'IMT

Plusieurs questions doivent être prises en compte en vue d'exploiter les Big Data pour l'IMT:

- la nécessité de disposer d'informations approfondies et précises sur l'offre et la demande sur le MT dans chaque pays;

- la nécessité d'une coopération entre les institutions pour l'échange de données, ainsi que la signature d'accords entre les propriétaires de données pour permettre un processus de collecte de données fiable et évolutive;

- la nécessité d'une utilisation conjointe des sources de données, telles que l'internet, pour recueillir les offres d'emploi, les enquêtes axées sur l'économie informelle/grise et les enquêtes sur la main-d'œuvre.

Le renforcement de l'utilisation des Big Data dans les pays en développement et en transition

L'utilisation de Big Data pour l'IMT serait particulièrement utile aux pays en développement et en transition afin de faciliter l'adéquation entre l'offre et la demande sur le MT, de concevoir des parcours professionnels répondant mieux aux attentes du MT et de comparer le MT interne à celui des pays transfrontaliers en vue de faciliter la mobilité. Dans ces pays, les aspects suivants doivent être soigneusement examinés:

- le manque d'accès aux données administratives et aux données statistiques;

- l'insuffisance des statistiques sur le système éducatif de ces pays;

- la nécessité d'un examen des sources internet pour évaluer la pénétration de l'utilisation d'internet pour les activités liées au MT;

- la nécessité d'une meilleure sensibilisation à l'importance des décisions fondées sur les données, tant pour le gouvernement que pour les offices de statistiques.

Des idées de projets pilotes

Compte tenu des commentaires recueillis lors de l'atelier précité («Skills needs uncovered: dealing with uncertainty») et de notre expérience en matière de Big Data pour l'IMT, nous identifions trois actions distinctes pour les projets qui pourraient aider les pays en développement et en transition à exploiter les Big Data pour le renseignement sur le MT:

- **Action 1.** Étude de faisabilité pour un pays X afin d'identifier, de valider et de classer les sources internet.

- **Action 2.** Mise en place d'un système de collecte d'informations en temps réel sur le MT.

- **Action 3.** Définition de modèles d'analyse des données pour aider les décideurs politiques dans l'élaboration et l'évaluation des politiques.

RÉFÉRENCES

[1] UK Department for Education and Skills, *LMI Matters!* 2004.

[2] UK Commission for Employment and Skills, *The Importance of LMI, available at https://goo.gl/TtRwvS*. 2015.

[3] M. Mezzanzanica and F. Mercorio, 'Big Data Enables Labor Market Intelligence', in *Encyclopedia of Big Data Technologies*, Springer International Publishing, 2018, pp. 1–11.

[4] ETF - The European Training Foundation, *Labour Market Information Systems*. 2017.

[6] UK Office for National Statistics, *NOMIS: UK National Online Manpower Information System*. 2014.

[7] E. Johnson, *Can big data save labor market information systems?* RTI Press policy brief no. PB-0010-1608, RTI Press, Research Triangle Park, NC, viewed, 2017.

[8] C. B. Frey and M. A. Osborne, 'The future of employment: How susceptible are jobs to computerisation?', *Technol. Forecast. Soc. Change*, vol. 114, no. Supplement C, pp. 254–280, 2017.

[9] UNECE - The United Nations Economic Commission for Europe, *Using Administrative and Secondary Sources for Official Statistics: A Handbook of Principles and Practices*. 2015.

[10] The Italian Ministry of Labour and Welfare, *Annual Report about the CO System, available at http://goo.gl/XdALYd*. 2012.

[11] S. Penneck and others, 'Using administrative data for statistical purposes', *Econ. LABOUR Mark. Rev.*, vol. 1, no. 10, p. 19, 2007.

[12] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, 'A model-based evaluation of Data quality activities in KDD', *Inf. Process. Manag.*, vol. 51, no. 2, pp. 144–166, 2015.

[13] R. Y. Wang and D. M. Strong, 'Beyond Accuracy: What Data Quality Means to Data Consumers', *J. Manag. Inf. Syst.*, vol. 12, no. 4, pp. 5–33, 1996.

[14] A. McAfee, E. Brynjolfsson, T. H. Davenport, D. Patil, and D. Barton, 'Big data: the management revolution', *Harv. Bus. Rev.*, vol. 90, no. 10, pp. 60–68, 2012.

[15] R. Boselli, M. Cesarini, F. Mercorio, and M. Mezzanzanica, 'Classifying online Job Advertisements through Machine Learning', *Future Gener. Comput. Syst.*, vol. 86, pp. 319–328, 2018.

[16] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, 'The KDD process for extracting useful knowledge from volumes of data', *Commun. ACM*, vol. 39, no. 11, pp. 27–34, 1996.

[17] T. C. Redman, 'The impact of poor data quality on the typical enterprise', *Commun. ACM*, vol. 41, no. 2, pp. 79–82, 1998.

[18] M. Bostock, V. Ogievetsky, and J. Heer, 'D^3 data-driven documents', *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011.

[19] F. Bao and J. Chen, 'Visual framework for big data in d3. js', in *Electronics, Computer and Applications, 2014 IEEE Workshop on*, 2014, pp. 47–50.

[20] CEDEFOP, *A new impetus for European cooperation in Vocational Education and Training to support the Europe 2020 strategy (Publicly available at https://goo.gl/Goluxo)*. European Commission, Strasbourg, France, 2010.

[21] European Commission, *A New Skills Agenda for Europe" COM(2016) 381/2*. European Commission, Strasbourg, France, 2016.

[22] EuroStat, *The ESSNet Big Data Project - available at https://goo.gl/EF6GtU*. European Commission, Strasbourg, France, 2016.

[23] CEDEFOP, *Real-time Labour Market information on skill requirements: feasibility study and working prototype". Cedefop Reference number AO/RPA/VKVET-NSOFRO/Real-time LMI/010/14. Contract notice 2014/S 141-252026 of 15/07/2014 https://goo.gl/qNjmrn*. 2014.

[24] CEDEFOP, *Real-time Labour Market information on Skill Requirements: Setting up the EU system for online vacancy analysis AO/DSL/VKVET-GRUSSO/Real-time LMI 2/009/16. Contract notice - 2016/S 134-240996 of 14/07/2016 https://goo.gl/5FZS3E*. 2016.

[25] UK Commission for Employment and Skills, *LMI4All - http://www.lmiforall.org.uk/*. 2015.

[26] M. Arntz, T. Gregory, and U. Zierahn, 'The risk of automation for jobs in OECD countries', 2016.

[27] - The Brookfield Institute for Innovation, *Better, Faster, Stronger: Maximizing the benefits of automation for Ontario's firms and people - https://brookfieldinstitute.ca/report/better-faster-stronger/*. 2018.

[28] T. A. Leopold, V. Ratcheva, and S. Sahiri, 'The future of jobs—Employment, skills and workforce strategy for the fourth industrial revolution, global challenge insight report', in *World Economic Forum, Geneva*, 2016.

[29] M. Mezzanzanica, F. Mercorio, and E. Colombo, 'Digitalisation and Automation: Insights from the', *Dev. Ski. Chang. World Work Concepts Meas. Data Appl. Reg. Local Labour Mark. Monit. Eur.*, p. 259, 2018.

[30] ESCoE - The Economic Statistic Centre of Excellence, *Using administrative and big data to improve labour market statistics - https://www.escoe.ac.uk/projects/using-administrative-big-data-improve-labour-market-statistics/*. 2019.

[31] M. Stonebraker, 'SQL databases v. NoSQL databases', *Commun. ACM*, vol. 53, no. 4, pp. 10–11, 2010.

[32] J. Dean and S. Ghemawat, 'MapReduce: simplified data processing on large clusters', *Commun. ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[33] E. Alpaydin, *Introduction to machine learning*. MIT press, 2009.

[34] E. Colombo, F. Mercorio, and M. Mezzanzanica, 'Applying machine learning tools on web vacancies for labour market and skill analysis', in *Terminator or the Jetsons? The Economics and Policy Implications of Artificial Intelligence*, 2018.

[35] F. Sebastiani, 'Machine learning in automated text categorization', *ACM Comput. Surv. CSUR*, vol. 34, no. 1, pp. 1–47, 2002.

[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, 'Latent dirichlet allocation', *J. Mach. Learn. Res.*, vol. 3, no. Jan, pp. 993–1022, 2003.

[37] R. Boselli *et al.*, 'WoLMIS: a labor market intelligence system for classifying web job vacancies', *J. Intell. Inf. Syst.*, vol. 51, no. 3, pp. 477–502, Dec. 2018.