Thèse
de doctorat
de l'UTT

Elamin ABDERRAHIM

Know-linking : favoriser un partage systématique et ciblé des connaissances entre les acteurs de l'entreprise

Champ disciplinaire :
Sciences pour l'Ingénieur

2022TROY0001 **Année 2022**

utt
UNIVERSITÉ DE TECHNOLOGIE
TROYES

THESE

pour l'obtention du grade de

DOCTEUR
de l'UNIVERSITE DE TECHNOLOGIE DE TROYES

en SCIENCES POUR L'INGENIEUR

Spécialité : SYSTEMES SOCIOTECHNIQUES

présentée et soutenue par

Elamin ABDERRAHIM

le 20 janvier 2022

Know-linking : favoriser un partage systématique et ciblé des connaissances entre les acteurs de l'entreprise

JURY

M. H. SNOUSSI	PROFESSEUR DES UNIVERSITES	Président
Mme M.-H. ABEL	PROFESSEURE DES UNIVERSITES	Rapporteure
Mme S. BRINGAY	PROFESSEURE DES UNIVERSITES	Rapporteure
Mme I. SAAD	ENSEIGNANTE-CHERCHEURE ESC AMIENS - HDR	Examinatrice
M. H. ATIFI	MAITRE DE CONFERENCES	Directeur de thèse
Mme N. MATTA	PROFESSEURE UTT - HDR	Directrice de thèse

Personnalité invitée

M. V. MAUGIS	RESPONSABLE DE LA GESTION DE CONNAISSANCES - ANDRA

“Avoir des objectifs est non seulement nécessaire pour nous motiver, mais c’est essentiel pour que nous restions en vie.” Robert H. Schuller

Remerciements

Je tiens tout d'abord à remercier les membres du jury qui m'ont fait le grand honneur d'accepter d'évaluer cette thèse : Madame Marie-hélène ABEL professeur des universités à l'UTC, Madame Sandra BRINGAY professeur des unviersités UPVM3, Madame Inès SAAD enseignant-chercheur (HdR) à ESC Amiens, Monsieur Hichem SNOUSSI professeur des universités à l'UTT, et Monsieur Vincent MAUGIS, responsable de la gestion des connaissances à l'agence nationale pour la gestion des déchets radioactifs. Je les remercie infiniment pour le temps consacré à cet effet en dépit de toutes les responsabilités qu'ils ont.

À l'issue de la rédaction de ce mémoire, je suis convaincu que la thèse est loin d'être un travail solitaire. En effet, je n'aurais jamais pu réaliser ce travail doctoral sans le soutien d'un grand nombre de personnes dont la générosité, la bonne humeur et l'intérêt manifestés à l'égard de ma recherche m'ont permis de progresser dans cette phase délicate de « l'apprenti- chercheur ».

Je tiens à remercier infiniment mes directeurs de thèse, Mme Nada Matta et M. Hassan Atifi pour la confiance qu'ils m'ont accordé en acceptant d'encadrer ce travail doctoral, pour leurs multiples conseils et pour toutes les heures qu'ils ont consacrées à diriger cette recherche. J'aimerais également leur dire à quel point j'ai apprécié leur grande disponibilité et leur respect sans faille des délais serrés de relecture des documents que je leur ai adressés. Enfin, j'ai été extrêmement sensible à leurs qualités humaines d'écoute et de compréhension tout au long de ce travail doctoral.

Mes remerciements vont également à l'école doctorale de l'université de technologie de Troyes, plus spécialement au directeur de l'école doctorale Monsieur Khemais Saanouni, Mme Pascale DENIS et Mme Isabelle LECLERCQ.

Je souhaiterais exprimer ma gratitude à mes parents, ma tante, mes frères, ma belle-soeur et mes cousins Ahmed, Aziz, Fatma et Jebril, qui ont toujours été à mes côtés.

Mes remerciements s'adressent aussi à tous mes amis et à mes collègues de l'UTT pour les moments agréables que nous avons passés ensemble et pour nos conversations enrichissantes.

Résumé

La connaissance est un capital qui a de la valeur pour l'entreprise. Ce capital peut être perdu pour plusieurs raisons, tel que le départ des experts qui ne laissent pas de traces, ou la perte d'accès à certaines ressources. Le besoin de réutiliser l'expérience passée et de partager des connaissances pour la réalisation des projets est devenu une exigence. L'objectif des entreprises depuis plusieurs années est alors de construire une stratégie réussie pour gérer et partager les connaissances, surtout dans un contexte qui a subi une évolution importante. Les méthodes traditionnelles ont montré leurs limites à l'ère de l'explosion des données et de l'évolution technologique. Dans ce contexte, nous proposons notre projet de thèse, fruit d'une étude bibliographique qui a conduit à l'élaboration de l'approche Know-linking, une approche composée de trois étapes dont la première consiste à profiler les collaborateurs de l'entreprise afin d'analyser leurs besoins en connaissances. Dans la deuxième étape, une représentation des profils est générée sous forme de graphes et les documents sont fouillés pour retrouver les liens sémantiques entre ces profils. La troisième étape est une étape de distribution de supports de connaissances selon les profils et liens entre eux. Notre approche favorise une génération de connaissances partagées d'une façon personnalisée (par profil de collaborateur) et systématique (assurée par un système). Nos travaux ont conduit à la construction d'une infrastructure de test de l'approche.

Mots clés :

- Échange de connaissances
- Traçabilité
- Profilage
- Traitement automatique du langage naturel
- Gestion des documents

Abstract

Knowledge is a valuable asset for the company, which can be lost for many reasons, such as the departure of experts who do not leave a trace, or the loss of access to certain resources. The need to reuse past experience and share knowledge for project implementation has become a requirement. The goal of companies for several years has been to build a successful strategy to manage and share knowledge, especially in a context that has undergone significant change. Traditional methods have shown their limits in the era of data explosion and technological evolution. In this context, we propose this thesis project, which is the result of a bibliographical study that led to the development of the Know-linking approach. This approach consists of three steps, the first of which consists of profiling the company's employees in order to analyze their knowledge needs. In the second step, we generate a representation of the profiles in the form of graphs and we search the documents to find the semantic links between these profiles. The third step is a distribution of knowledge supports according to the profiles and links between them. Our approach favors the generation of shared knowledge in a personalized (by employee profile) and systematic (provided by asystem) way. Our work led to the development of a framework for the approach.

Keywords :

— Knowledge sharing
— Traceability
— Profiling
— Natural language processing
— Records-Management

Table des matières

Table des figures

Liste des tableaux

à l'âme de ma grand-mère .....

Première partie

Introduction, contexte et problématique

Chapitre 1

Introduction

"Le commencement est la moitié de tout." Pythagore

Sommaire

1.1 Contexte de recherche

La connaissance représente le centre d'intérêt de plusieurs travaux de recherche qui visent à la définir, à la contextualiser et à la comprendre comme un phénomène garant de la continuité du développement humain. Peirce, par exemple, a présenté la théorie du triangle sémiotique basée sur la théorie sémiotique de Saussure qui distinguait le signifiant et le signifié [Eco 1979]. D'après la théorie du triangle sémiotique, trois centres de triangles décrivent trois différentes dimensions [Chandler 2007]. La connaissance, d'après cette théorie sémiotique, est un symbole qui a un sens en se basant sur une référence. On distingue

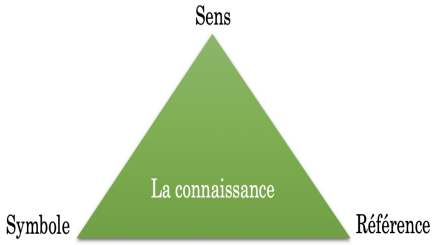


FIGURE 1.1 – Triangle sémiotique

deux familles de connaissances qui sont respectivement les connaissances tacites et les connaissances explicites [Polanyi 1961]. Le mot tacite caractérise les connaissances non formalisées, non décrites et qui résident dans la mémoire d'un être humain ou dans son imagination. Les connaissances explicites, quant à elles, sont des connaissances décrites, formalisées ou écrites [Polanyi 1961]. Une connaissance tacite peut être transformée en explicite et vice-versa d'après le modèle SECI présenté par Nonaka [Nonaka *et al.* 1995].
Ce modèle repose principalement sur quatre types de conversion :

- L'externalisation : L'expression de connaissances tacites et leur traduction en des formes compréhensibles par d'autres. L'individu dépasse les barrières internes et externes du soi, il s'engage dans un groupe et se transforme en une seule unité ;
- La socialisation : La connaissance tacite est partagée à travers des activités communes. Le rapprochement physique est une condition nécessaire à ce type de transmission des connaissances ;
- L'internalisation : processus de création de nouvelles connaissances tacites à partir de connaissances explicites ;

— La combinaison : La mise en commun de plusieurs connaissances explicites afin de les mettre à disposition du collectif.

Étant donné que le partage de la connaissance impacte étroitement son accessibilité, Nonaka a étudié le contexte dans lequel la connaissance peut être créée et convertie. Il a présenté le concept du 'BA', inspiré des travaux du philosophe Kitaro Nishida, comme un idéogramme kanji composé de deux parties gauche et droite [Nonaka & Konno 1998]. La partie gauche peut être assimilée à la terre, à l'eau bouillante ou à ce qui soulève ; la partie droite quant à elle signifie ce qui rend possible (enable) [Fayard 2002]. Le BA est une traduction japonaise du mot lieu ou emplacement, autrement dit cet espace partagé pour les relations émergentes est un contexte qui sert de base à la création de connaissances [Fayard 2002].

La connaissance est intégrée dans le BA et elle est acquise soit par l'expérience de l'individu, sa réflexion, soit à travers l'expérience des autres. Le BA peut être physique, virtuel, mental ou une combinaison de tous.

Cette définition du concept du BA nous conduit à considérer l'entreprise, ce large espace émergent de connaissances, comme un BA physique dans la mesure où l'entreprise est un emplacement de création et de génération de la connaissance. En effet, dans une entreprise, un nombre important de collaborateurs de différentes spécialités interagissent autour de processus de production complexes et intersectés [Bessire & Mesure 2009]. En réalité, cette collaboration suscite de très nombreux supports de connaissances de différents types et formats (systèmes ou documents) sous forme de justifications des choix, des études, des spécifications, des propositions de concepts ou des mesures scientifiques.

La connaissance éparpillée sur ces supports est produite en interaction tout au long de la réalisation de l'activité. Elle est fondamentale pour la continuité et la survie de l'entreprise, autrement dit les collaborateurs ont toujours besoin de la connaissance intégrée dans l'entreprise pour leur travail quotidien. Cette connaissance est dite partagée car l'ensemble des collaborateurs de l'entreprise partagent le besoin d'y accéder, un besoin qui est exprimé au fil du temps. Cette connaissance est un capital qui est estimé à 217% du capital financier net de l'entreprise d'après une étude faite par le "knowledge strategist" Paul strassman sur un ensemble d'entreprises aux USA [1].

Gérer les connaissances d'une organisation est une discipline qui englobe la création, le partage ainsi que la capitalisation de ce capital [Grundstein 2009] [Ermine 2003]. Face au développement galopant des technologies, des outils et des méthodes dans l'industrie, les méthodes de gestion de connaissance classique se

1. Plusieurs études effectuées par Paul Strassman,https ://www.strassmann.com/articles.html ?t=all

retrouvent timides face à ce changement. En se basant sur la théorie de l'iceberg qui définit l'iceberg comme un amas de glace dont 10 à 20% de sa masse est visible contre 80 voire 90% invisible, on peut affirmer que les entreprises peuvent être perçues comme un iceberg dont les 10% visibles sont les chiffres et les produits et que la complexité et les problèmes se cachent dans les 90% de sa partie invisible (figure 1.2).

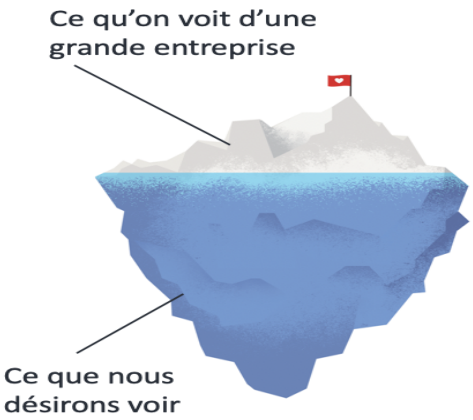


FIGURE 1.2 – Iceberg de l'entreprise

Le problème majeur de l'entreprise d'aujourd'hui réside au niveau du partage des connaissances [NINTEX]. **«Qu'avez-vous comme problème ?»** une question qui a été posée à 1000 employés dans une enquête menée par NINTEX [NINTEX] dans des entreprises aux États-unis, 49% de ces employés ont affirmé qu'ils souffrent d'un problème d'accès aux documents ! (figure **??**)

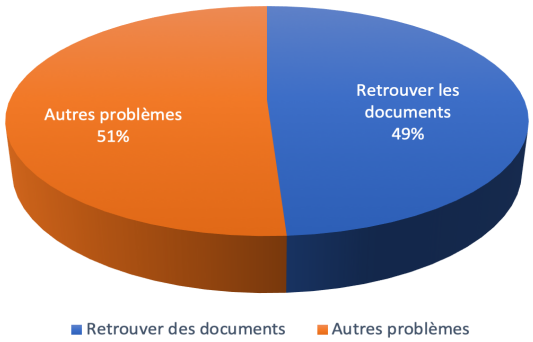


FIGURE 1.3 – Les problèmes des employés dans l'industrie

Le phénomène du partage de la connaissance dépasse le cadre de l'entreprise. Depuis l'Antiquité, l'être humain essaye de transmettre son savoir et son savoir-faire par le biais des écritures hiéroglyphiques chez les pharaons par exemple ou des mosaïques carthaginoises. Des symboles représentés sur des murs ou sur des tableaux formalisant le savoir-faire de ces civilisations de l'Antiquité en agriculture, médecine et autres domaines. Pour pouvoir partager la signification de ces symboles il faut tout d'abord les retrouver et les déchiffrer ! Dans l'entreprise, partager la connaissance c'est la rendre accessible à l'ensemble des parties prenantes qui en éprouvent le besoin. L'organisation de certaines entreprises, notamment les grands groupes, représente un obstacle face au partage des connaissances. Le travail en silo, où chaque employé produit des connaissances, représente un obstacle au partage des connaissances entre collègues. Le manque d'échange entre les employés, l'absence d'information sur la production de certaines connaissances doivent être évités pour l'accomplissement d'une activité. L'employé se retrouve isolé dans son environnement technique !
Nous pouvons donc simuler ce problème comme suit : chaque employé isolé dans son environnement se croit indépendant, ne se rend pas compte de l'importance de la collaboration avec ses collègues et souffre de problèmes d'accès ainsi que de l'ignorance de l'existence de supports de connaissances dont il a besoin.

1.2 Problématique de recherche

L'étude des points critiques dans l'entreprise que nous avons observés et présentés dans la section précédente nous a aidés à identifier des problèmes liés au partage de connaissances dans les organisations. Nous détaillerons ces problèmes dans la partie suivante.

1.2.1 Connaissances éparpillées

Dans un contexte industriel où les entreprises suivent l'évolution galopante des outils et des méthodes de travail, la production des sources de connaissances a par conséquent évolué d'une façon considérable. Les connaissances de l'entreprise sont éparpillées dans les documents électroniques (excel, word, pdf..), dans les emails, les systèmes d'informations (outils de gestion de projet, outils de gestion des ressources humaines).
D'après certaines études, les documents représentent 20% du total des sources de connaissances et parmi eux 26% constituent des supports papier face à 42% de connaissances tacites (portées par les collaborateurs) [Segonds 2011].

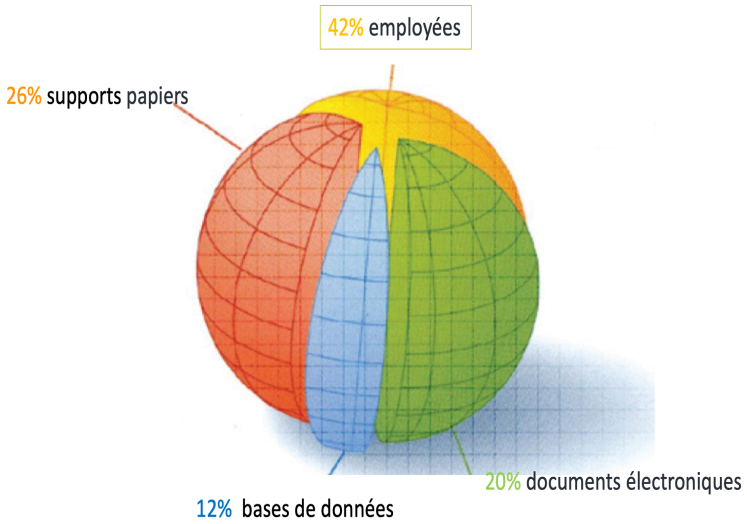


FIGURE 1.4 – Supports de connaissances [Segonds 2011]

Les supports de documents (électroniques ou papiers) représentent un pourcentage important de la totalité des sources d'après cette étude, et ce pourcentage n'a pas cessé d'augmenter depuis 2011 jusqu'aujourd'hui suite à la croissance du contenu numérique.
La connaissance dans un tel environnement est aussi distribuée. Les connaissances distribuées peuvent être considérées comme la somme des connaissances dans un groupe. On l'appelle parfois la connaissance potentielle d'un groupe, ou la connaissance commune qu'ils pourraient obtenir s'ils disposaient de moyens de communication illimités [A & Wáng 2017]. Pour conclure, l'absence d'une stratégie de centralisation du savoir-faire de l'entreprise a rendu la connaissance distribuée et éparpillée sur une multitude de supports.

1.2.2 Connaissances peu accessibles

Les études conduites par NINTEX dans son rapport « Definitive guide to America's Most Broken Processes » ont démontré que 55% des échecs d'intégration des nouveaux employés dans l'entreprise sont causés par le problème d'accès aux documents et aux outils. De même, 49% de tous les employés de l'entreprise rencontrent des problèmes de localisation des documents [NINTEX]. La connaissance intégrée dans ces supports est peu accessible. Les collaborateurs perdent donc du temps à rechercher des informations et à dupliquer les études faites par leurs collègues.

1.2.3 Connaissances hétérogènes et diversifiées

La complexité des projets nécessite une expertise dans différents domaines. Par exemple, pour un projet d'aéronautique, plusieurs savoirs sont mobilisés en électronique, mécanique, aérodynamique, etc. Ceci engendre une diversité de connaissances à la fois requises et produites.

1.2.4 Manque de traçabilité

Les approches de traçabilité peuvent résister à une fuite des connaissances prévue dans l'entreprise. La traçabilité permet de garder une trace de la mémoire épisodique dans laquelle des associations spatio-temporelles des évènements sont décrites [Karsenty *et al.* 2001]. Plusieurs approches ont été développées dans l'industrie pour garder les traces de la connaissance, certaines d'entre elles sont basées sur la structuration des entretiens avec les experts dans des fiches comme REX [Malvache & Prieur 1993], ou les fiches MEREX [Corbel 1997] pour sauvegarder les bonnes et les mauvaises expériences. D'autres sont basées sur la structuration des PV des réunions de manière à faire la différence entre proposition, décision et critère pour la mémoire du projet [Matta *et al.* 2013b]. Cependant, la plupart des entreprises n'adoptent pas de solution efficace pour garder la trace de la connaissance.

1.2.5 Formulation de la problématique de recherche générale

Les différents points présentés dans la section précédentes sont le moteur d'un réel problème stratégique dans l'entreprise qui consiste principalement à :

- **Un risque sérieux de perte des connaissances :** une connaissance peut être facilement perdue en cas d'absence d'une approche efficace de traçabilité et de capitalisation.
- **La redondance :** vu que l'accès aux connaissances éparpillées et distribuées est une tâche ardue qui nécessite en plus des efforts de recherche, un effort supplémentaire d'enquête et d'investigation est demandé aux acteurs d'une entreprise. Les collaborateurs produisent des études et des documents qui existent déjà.
- **Un bruit :** la connaissance demandée est, dans la plupart des cas, distribuée sur plusieurs supports. Un collaborateur a du mal à l'identifier.
- **Un temps de la recherche de connaissances élevé :** en l'absence d'une solution pertinente et de réponses aux requêtes des collaborateurs,

le temps requis pour fouiller toute une masse volumineuse de supports hétérogènes est important.

La problématique que nous abordons dans ce projet de thèse est donc, **la difficulté de partager les connaissances d'une façon efficace garantissant l'accès rapide et facile de chaque collaborateur aux connaissances dont il a besoin**.

1.2.6 Questions de recherches

Mettre en place une stratégie de partage des connaissances de l'entreprise constitue l'axe de nos recherches. L'étude de la problématique en profondeur a suscité quelques réflexions traduites sous formes de questions de recherches auxquelles nous essayons de répondre :

1. **Comment générer la connaissance partagée entre les différents acteurs d'une façon régulière et dynamique ?**
   Le collaborateur est l'axe central de l'entreprise, c'est lui qui crée la connaissance. En analysant la structure organisationnelle de l'entreprise par les techniques du profilage, nous pouvons représenter les productions des collaborateurs dans des structures appelées «profils de collaborateurs». Ces profils permettent de faciliter l'identification des connaissances que les collaborateurs produisent et dont ils ont besoin au fil du temps. La génération des graphes sémantiques de ces profils, (à l'aide du traitement de langage naturel [Baclic *et al.* 2020]) facilite la liaison d'un côté du profil avec son environnement de connaissances, et d'un autre côté le profil avec les autres profils générés, ce qui dévoile les collaborations indirectes qui demandent des connaissances partagées. D'après nous, deux facteurs sont à prendre en considération pour pouvoir générer les connaissances partagées : le rapport du profil avec l'environnement de connaissance de l'entreprise et avec les autres profils de collaborateurs.
   La mise en place d'un système basé sur ce principe assurera l'enrichissement des graphes des profils par de nouveaux concepts extraits d'analyses textuelles de documents produits par les collaborateurs.
2. **Comment assurer une traçabilité efficace des connaissances ?**
   La traçabilité permet de garder une trace de la mémoire épisodique dans laquelle des associations spatio-temporelles des évènements sont décrites [Karsenty *et al.* 2001]. Nous proposons une traçabilité personnalisée par la construction d'une structure de profil complète considérant la

définition de l'ensemble des tâches qu'un collaborateur réalise, les projets dans lesquels il est impliqué ainsi que les outils qu'il utilise.

3. **Comment réduire le bruit ?**
   La construction des graphes de profils nous offre une représentation de la production de l'ensemble des collaborateurs et des interactions entre eux. Certaines interactions sont évidentes, bien exprimées dès le départ, tandis que d'autres restent cachées dans les documents, construites tout au long des projets. L'analyse du contenu textuel avec le traitement automatique des langages naturels [Guillén *et al.* 2017], peut dévoiler ces interactions cachées derrière des liens linguistiques qui reflètent des liens sémantiques entre les profils. La détection de ces liens diminue d'une façon considérable le bruit des connaissances. Dans la mesure où la représentation de chaque profil va être reliée aux documents dont un collaborateur a besoin par l'indexation et va fournir un accès aux documents produits par ses collègues qui entrent dans son optique d'intérêt.

1.3 Hypothèse de solution

Pour générer la connaissance de l'entreprise d'une façon régulière et dynamique, nous proposons donc l'approche Know-linking qui repose sur trois principes de base :

— **Génération des profils de collaborateurs de l'entreprise :**
  Une étape de structuration de l'organisation des collaborations de l'entreprise sous forme de profils. Chaque profil est identifié par une liste de missions, d'outils et de vocabulaires ;

— **Génération des graphes des profils et de liens sémantiques entre les profils :**
  Il s'agit d'identifier une représentation sémantique des profils de collaborateurs en révélant les liens sémantiques cachés dans les documents ;

— **Distributions des documents selon les graphes des profils :**
  Pour chaque profil nous identifions les documents liés à sa spécialité et en rapport avec son métier ainsi que les documents qu'il partage avec d'autres profils de collaborateurs.

L'approche Know-linking est composée principalement de trois étapes, modélisée comme suit, dans la figure 1.5.

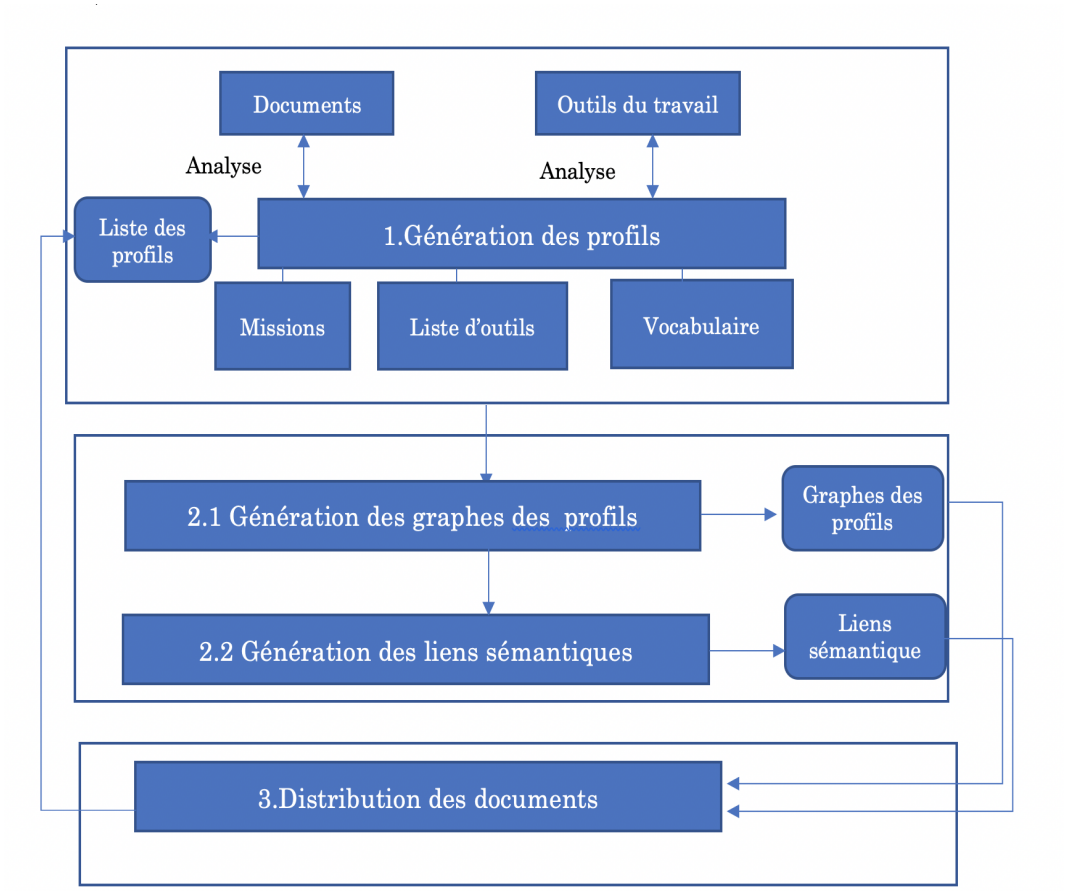


FIGURE 1.5 – L'approche Know-linking

1.4 Organisation du manuscrit

Dans ce rapport de thèse nous présentons une approche de génération dynamique et régulière de connaissances basée sur les profils de collaborateurs. Le rapport de la thèse est structuré suivant quatre modules. Le premier module concerne la présentation du contexte et du cadre général de la thèse, le second est dédié à l'état de l'art. Quant au troisième module, il est consacré à la présentation de notre approche 'Know-linking'. Dans la dernière partie nous concluons le manuscrit par une discussion de nos travaux tout en présentant des perspectives de travaux afin de combler certaines limites de notre approche. Nous détaillons cette structure dans la figure 1.6 suivante.

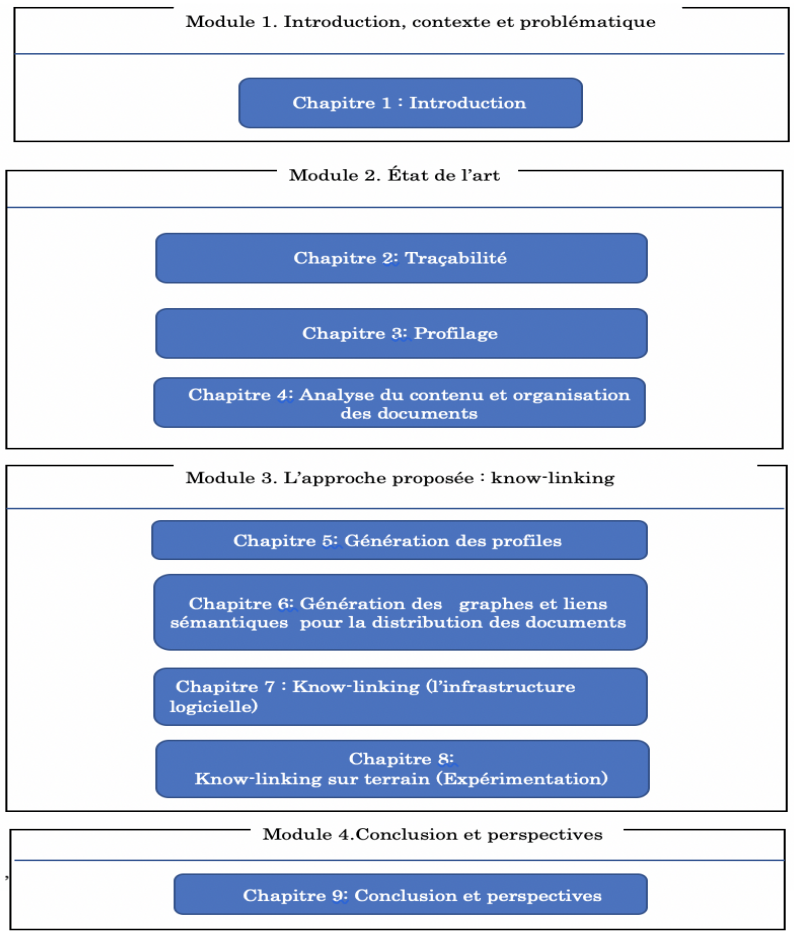


FIGURE 1.6 – Organisation du manuscrit

Deuxième partie

État de l'art

Chapitre 2

La Traçabilité

« **Un Homme qui ne marche pas ne laisse pas de traces** » Georges Wolinski

Sommaire

2.1 Introduction

Une trace de projet est constituée à partir d'empreintes laissées volontairement ou non dans l'environnement à l'occasion d'un processus [Mille 2013]. Ces empreintes seront relevées comme traces de passage d'un objet ou d'un être. Ces traces permettent d'identifier le chemin parcouru. Dans l'entreprise, une trace peut se présenter comme un support à la mémoire. Elle peut prendre la forme d'un document, d'un support électronique ou même d'un email.
Dans le présent chapitre nous présentons les différents travaux de définition de la trace et nous expliquons ce que nous voulons dire par « trace » et « traçabilité ». Nous discutons les approches existantes afin d'identifier les traces d'un projet et nous analysons de quelle manière ces traces collectées et sauvegardées participent éventuellement à la construction d'une mémoire organisationnelle de l'entreprise. La traçabilité présente un lien entre le présent de l'entreprise et son passé. Ce lien permet de revenir en arrière et d'en profiter, ce qui va nous aider dans le contexte de notre thèse, de remédier à la rupture de l'entreprise avec son passé causé par la mobilité des experts et le manque de la sauvegarde de la trace. Ce travail de bibliographie permet de comprendre et d'analyser les travaux existants sur la traçabilité et la mémoire du projet et ce afin de répondre à un axe central de notre problématique de recherche qui est le besoin de la réutilisation de l'expérience passée de l'entreprise. Dans cette perspective, nous consacrons une section à la fin du chapitre afin de le détailler.

2.2 Trace et traçabilité

Le terme trace rappelle les traces de pas qu'un animal ou qu'un être humain laisse dans le sable ou encore dans la neige après son

passage comme le montre la figure 2.1 ci dessous.

FIGURE 2.1 – Traces animales

Au-delà de cette image, ce terme a pu acquérir un sens polysémique. À la base, ce terme provient du verbe tracer qui tient son origine du latin « tractiare » et qui signifie l'action de tracer ou de tirer ainsi que la lenteur [Laflaquière 2009].
Selon le dictionnaire Petit robert (2005), « une trace est une chose ou une suite de choses laissées par une action quelconque et relatives à un être ou un objet ». Une trace peut être également considérée comme une « empreinte » d'après le même dictionnaire Petit robert : « une suite d'empreintes ou de marques que laisse le passage d'un être ou d'un objet ». Grâce à cette trace nous pourrions reconnaitre l'existence d'un objet ou d'un être.
Dans certains travaux une trace est définie comme le partage des caractéristiques avec la notion d'indices dans la mesure où une trace peut disparaitre ou échapper au regard d'observateurs tout comme un indice [Laflaquière 2009]. La caractérisation d'une trace dépend alors de l'observation de l'observateur.
D'autres visions rapprochent le terme trace avec la notion du signe. En effet, les deux notions « trace » et « signe » étaient indifféremment utilisées par le passé [Demonet-Launay 1994]. Le signe, d'après Augustin, est « ce qui, se présente en tant que tel à la

perception sensible, présente aussi quelque chose à la perception intellectuelle » [Demonet-Launay 1994].

La trace ne peut pas être définie par elle-même car, réellement, elle n'a pas d'existence propre ou autonome au plan ontologique du terme. Elle n'existe que par rapport à une entité, par exemple un événement, un être ou un phénomène [Serres 2002].

En médecine, la trace est l'altération postulée dans les cellules du système nerveux qui se produit en tant que résultat de toute expérience ou apprentissage. Dans la conception, une trace est définie par un triple spécifié d'éléments comprenant : un artefact source, un artefact cible et une liaison (dite de trace) associant les deux objets.

L'utilisation du terme trace a connu une évolution au fil du temps. Par exemple, au 17ème siècle, le terme tracer est utilisé pour représenter au moyen de lignes ou « marquer le contour ». Quelques siècles plus tard, le même terme est utilisé dans un « paradigme indiciaire » [Ginzburg 1989]. Au 20ème siècle, le terme est utilisé dans un contexte informatisé donnant par conséquent la naissance au terme « trace numérique ». Nous nous intéressons dans nos travaux à cette trace numérique de l'activité humaine dans l'entreprise, notamment la trace des connaissances générées tout au long de cette activité. Dans ce cadre, la trace est considérée comme une modification plus ou moins durable de l'environnement [Rauscher 2016]. Ces traces sont le produit d'une action volontaire ou involontaire. Selon [Rauscher 2016] les traces gardent un enregistrement des interactions entre les collaborateurs dans une entreprise [Rauscher 2016] tels que les échanges par emails, à travers les réseaux sociaux, ...

Le dictionnaire Larousse considère la traçabilité comme une « possibilité de suivre un produit aux différents stades de sa production, de sa transformation et de sa commercialisation, notamment dans

les filières alimentaires. » Cette définition met l'accent sur la notion de 'suivi' du projet à différentes étapes. Implicitement, on peut lier les traces à toutes les phases de la réalisation du projet. Un tel lien est maintenu après une observation du processus métier où plusieurs acteurs interviennent à des moments différents et pour des raisons différentes. La traçabilité c'est sauvegarder les liens entre le qui (le participant), le quand (le temps), le quoi (la raison) en rapport avec le comment (l'activité). Ces éléments sont ainsi considérés comme des traces de l'activité. [Lund & Mille 2009] considèrent par exemple dans leurs travaux que la collecte de telles séquences d'observées temporellement situées fournit les « traces » de l'activité d'apprentissage qui seront la source de connaissances pour le processus de personnalisation quel que soit son niveau. La collecte de ces traces dans un environnement spécifique permet d'avoir un espace de mémoire partagé que Louise Merzeau [Merzeau 2013] appelle à dépasser pour que *«la traçabilité ne relève plus seulement d'une indexation plus ou moins maîtrisée de soi, mais d'une construction d'espaces communs de connaissance et de mémoire »*. La traçabilité est simplement « le potentiel de ces traces à être établi (créé et maintenu) et utilisé ». La traçabilité ou « garder trace » est l'action de suivre des traces afin d'identifier l'impact et l'évolution des événements sur l'environnement [Matta *et al.* 2013a].

Dans le contexte d'une large entreprise où les traces sont sauvegardées dans des supports et majoritairement dans des documents de différents formats, nous pouvons définir la traçabilité comme un lien temporel entre le collaborateur et son activité dans le but de réduire le gap sémantique entre l'ensemble des supports numériques (contenant les traces) et l'usager.

2.3 L'apport de la traçabilité de l'activité dans une entreprise

[Ramesh & Edwards 1993] cite l'exemple d'une entreprise qui a été obligée de réembaucher des experts qui ont quitté cette même entreprise par manque de traçabilité des activités et des logiques de décisions [Ramesh & Edwards 1993]. Les avantages qu'apportent la mise en place d'une approche de traçabilité dans une entreprise ne sont pas seulement stratégiques mais se veulent également financiers.

En effet, garder la trace de la connaissance produite dans une activité permet à une entreprise de :

- Avoir des réponses sur les justifications des choix de l'entreprise à un moment donné ;
- Avoir une vision sur les facteurs de création de la connaissance à un moment donné ;
- Avoir la possibilité de reproduire des études (tant qu'on connaît les démarches sous-jacentes) ;
- Profiter de l'expérience passée de l'entreprise et exploiter des connaissances produites dans un contexte temporel différent ;
- Assurer un partage de l'expertise à une échelle de temps plus large.

L'apprentissage, que la traçabilité peut garantir pour une entreprise, est conditionné non seulement par la bonne adaptation d'une approche de traçabilité efficace et pertinente mais également par l'application continue de cette approche avec des systèmes.

2.4 Les Systèmes traçants

Les systèmes dits « traçants », s'ils permettent d'enregistrer les traces « dans un but précis », ne sont pas le fait du « simple fonctionnement des logiciels utilisés» [Laflaquière 2009]. Il existe une variété d'outils modifiés ou adaptés pour la collecte, la génération et le partage de la trace d'une façon systématique.
Dans cette section, nous présentons une étude bibliographique autour des systèmes et des approches permettant la traçabilité.

2.4.1 Les approches de traçabilité

D'un point de vue historique, les approches de traçabilité ont connu une évolution. Elles sont passées de simples approches de traçabilité de la nourriture pour l'hygiène à des approches plus complexe intégrées dans l'environnement de travail de l'entreprise pour capter les traces des experts et les sauvegarder. Dans notre étude nous nous intéressons à la traçabilité numérique. Dans ce qui suit, nous analysons ces techniques selon leurs principes :

— Portés plus sur la traçabilité de résolution de problèmes comme avec des entretiens des acteurs ou mémoriser la réalisation de projets ;
— Traçabilité des actions des utilisateurs de systèmes comme les logs des usages des logiciels ou l'exploitation des systèmes d'information pour la génération des connaissances.

2.4.1.1 Les approches basées sur les entretiens avec les experts

Toutes les traces ne sont pas « écrites » ou « formalisées », il existe en effet des traces non écrites qui sont appelées par Bloch [Damien 2012] [Serres 2002]« vestiges du passé » ou encore « témoignage non écrit ». Ce qui motive la définition des approches

sollicitant les acteurs du projet est leur implication dans ce processus de réflexion de l'entreprise en formalisant aussi l'ensemble des traces non écrites.

Le besoin d'une approche de traçabilité chez le commissariat à l'énergie atomique français, un exemple d'entreprise à large envergure, a été traduit par la mise en place de l'approche REX (retour d'expérience) [Malvache & Prieur 1993], le principe étant d'organiser des entretiens avec les experts de l'entreprise, pouvant être individuels ou collectifs, et de structurer par la suite ces entretiens dans des fiches appelées 'REX'. Ces fiches REX s'inscrivent dans un processus de réflexion de l'entreprise. L'approche REX est décrite de différentes manières dans la littérature scientifique. On peut trouver deux différents REX : un REX Métier et un REX Projet [Prax 2012]. Dans tous les cas d'application il s'agit de formaliser un descriptif des différentes composantes de l'action : ses caractéristiques techniques, ses parties prenantes, les participants.. etc. La mise en place de l'approche REX peut être assurée grâce à un expert, et/ou faire appel à un logiciel pour l'analyse et l'organisation d'une base de données [Prax 2012].

Renault à son tour a présenté sa propre solution de traçabilité pour la capitalisation des activités de conception, une approche basée sur des fiches MEREX (Mise En Règle de l'Expérience) [Corbel 1997]. Après un entretien avec l'expert on procède à l'illustration des bonnes ainsi que des mauvaises expériences. On désigne par les bonnes expériences « l'innovation » et par les mauvaises l'ensemble « des problèmes » et « solutions ». Les fiches MEREX sont composées de « check-list » permettant de contrôler un processus ou seulement une étape d'un processus. L'ensemble des documents produits à l'issu des entretiens est accrédité par les autres acteurs du processus concerné pour être enfin distribués aux acteurs [Prax 2012].

Bien que ces approches aient prouvé leur efficacité par leur application dans de grandes entreprises comme Renault ou CEA, les experts ont du mal à expliciter leurs travaux au quotidien, ce qui exige un effort supplémentaire de structuration des détails recueillis. Elles sont aussi coûteuses, et une forte implication d'un expert de Gestion de connaissances ou de traçabilité est demandée pour leur application.

2.4.1.2 Mémoire de projet

Un projet est défini comme « un ensemble fini comportant un début et une fin, un caractère unique, une aventure mêlant des expériences positives et négatives ». En fait, la réalisation d'un projet s'inscrit dans un processus mental dans la mesure où le projet est *« un lieu de co-construction de connaissances nouvelles, avec une structure bien définie»* [Rauscher 2016], ce qui sollicite des activités d'apprentissage et de résolution de problèmes. Récupérer cette dimension prend en compte les liens entre ce qui a été réalisé, comment il a été réalisé et surtout comprendre le pourquoi des décisions tout en gardant les alternatives non considérées dans la prise de décision. La Mémoire de Projet se focalise effectivement sur la conservation de « la définition du projet, les activités, l'historique et le résultat » [Tourtier 1995]. Pour comprendre ce que « l'historique du projet » peut inclure il a fallu attendre d'autres travaux pour mieux définir la mémoire du projet. D'après [Matta *et al.* 2013b] pour le citer comme exemple, la mémoire du projet englobe :

— L'organisation du projet : Les participants, leurs compétences, l'organisation de l'équipe, les tâches, etc ;
— Les cadres de référence : les règles et les réglementations ;
— La réalisation du projet : la résolution du problème et la gestion des risques ;

— Le processus de prise de décision : les négociations stratégiques et les résultats.

La récupération de la mémoire du projet avec toutes les dimensions qu'on vient de citer nécessite principalement deux travaux qui sont la collecte et la modélisation. Dans DYPKM [Bekhti 2003] par exemple, la collecte et la modélisation des connaissances en lien avec la conception d'un projet remédie au problème de la gestion de connaissances classique qui ignore le contexte d'un projet.

DYPKM propose un processus visant à définir un modèle relationnel global regroupant des éléments de contexte et de logique de conception. Les informations et le contenu généré lors de la réalisation du projet sont capturés au fur et à mesure puis structurés comme traces de projet. Les échanges formels dans la réalisation des projets forment un support de mémoire organisationnelle par exemple dans le cadre des réunions, des échanges multiples sur le budget, les choix, les propositions ainsi que les décisions, tout en se basant sur les facteurs de la réalité terrain (au moment de la réunion). Matta dans [Matta *et al.* 2013b] stipule que les éléments abordés par des participants dans des réunions comme les questions et les décisions peuvent être récupérés pour garder une trace du contexte de prise de décision de l'entreprise. Dans ce cadre s'inscrit le système 'Memory meeting' [Matta *et al.* 2013b] permettant de garder une trace des participants, des questions et des décisions dans les réunions organisées à cet effet. Les propositions et les arguments présentés lors de la réunion permettent de comprendre le contexte du processus de prise de décision considérés également dans cette mémoire des réunions.

Dans le cadre de la mémoire organisationnelle s'inscrit le projet MEMORAe [Abel *et al.* 2002] qui se base sur une modélisation ontologique de l'apprentissage (e-learning), pour appliquer une démarche d'ingénierie de connaissance dans un contexte éducatif.

Pour chaque domaine spécifique les auteurs proposent d'élaborer des ontologies qui permettent de lier les traces d'information et les participants par des annotations, autrement dit lier les « traces » à «l'organisation».

2.4.1.3 Approches basées sur l'analyse du log :

Les fichiers logs générés par les programmes informatiques lors de l'exécution peuvent être utilisés comme un système pour la traçabilité. Comme nous l'avons précisé précédemment, la traçabilité des systèmes d'information est définie comme la capacité à tracer les relations entre les différents artefacts du système. En effet, un log est un message texte avec des métadonnées contenant des informations sur un évènement produit au sein du programme [Ghania & Nora 2018].
Nous pouvons citer MUSETTE [Champin *et al.* 2004] comme exemple d'approche qui exploite les logs comme supports de traçabilité.
La finalité de cette approche est de construire des bases d'expériences d'utilisateurs à partir les liens établis entre les traces, les objectifs des utilisateurs et leurs activités. Pour guider l'utilisateur dans son travail, un système de raisonnement basé sur l'expérience *«Experience Based reasoning system »* permet d'exploiter la base d'expériences fournies. Un système de signatures des traces permet de reconnaître une expérience de la base afin de guider un acteur dans son activité.
Les logs peuvent aussi servir à l'analyse des traces d'interactions dans un environnement virtuel. Orklin, dans son approche, [Orkin & Roy 2010] se base sur l'analyse de 100 logs du jeu The restaurant contenant les interactions, notamment des dialogues entre les agents du système et les utilisateurs. Les logs sont annotés sous forme de triplets d'acte de dialogues (contenu, réfèrent et acte de parole). Cette analyse de traces d'interaction dans les logs a servi

à la construction d'un modèle d'apprentissage afin de prédire des dialogues dynamiques.
L'analyse des logs peut aussi mettre en exergue les liens qui existaient entre le code source d'un logiciel et les exigences pour la traçabilité des systèmes d'information. Ce lien est généralement découvert à l'aide des « similarity based method» [Tsuchiya *et al.* 2013]. Les traces de ces liens pouvaient être récupérées à partir des fichiers logs, ce qui poussait Ryosuke et al. à compléter le travail de traçabilité des exigences présenté dans leur approche [Tsuchiya *et al.* 2013] en 2013 par l'analyse des logs de gestion de configuration générés par des systèmes de gestion de version git et Subversion en 2015 [Tsuchiya *et al.* 2015].
Les fichiers logs contiennent aussi les données sur les bugs du programme informatique ce qui a poussé [Romo & Capiluppi 2015] à exploiter ce contenu pour réduire le gap entre les logs du développement et les données issues des bugs. La traçabilité des bugs dans ce travail permet de prédire les problèmes dans un système d'information. Les fichiers logs, bien que décrivant le comportement de l'utilisateur du système, sont généralement stockés dans une mémoire non permanente. Ce sont aussi des fichiers qui occupent une partie importante de l'espace mémoire les rendant ainsi facilement supprimables et perdus de la mémoire. Un autre point pouvant impacter la performance de ces approches est le contenu des logs noté dans un langage de programmation le rendant incompréhensible par le non-connaissant du même langage.

2.4.1.4 La traçabilité intégrée dans des systèmes d'information

Dans une logique de garder la trace d'informations d'une façon quotidienne, certains systèmes ont endance à intégrer un module assurant la traçabilité « en background », ce qui signifie que ces systèmes ne sont à la base pas dédiés à la traçabilité mais peuvent

être exploités pour capter « la connaissance quotidienne ».
La connaissance quotidienne, ou « daily knowledge », est une connaissance produite par l'être humain dans son quotidien [Matta *et al.* 2016]. Elle peut être considérée comme une mémoire épisodique qui participe à la construction de la connaissance épistémique [Richard 1992]. La capture de cette connaissance doit avoir un caractère non « intentionnel ». Cependant, la question qui se pose est la suivante : quel système d'information choisir pour cette mission ?
Chaque système d'information traite un point très particulier dans la réalisation du projet ce qui pose des problèmes d'interopérabilité entre les environnements au cas où on chercherait à lier tous les éléments du projet, si bien que la solution considérée doit couvrir toutes les phases de la réalisation d'un projet comme le Project Lifecycle Management(PLM). Le PLM est une approche stratégique pour la création et le management d'un capital intellectuel d'une organisation de la conception au retrait [Subrahmanian *et al.* 2005]. Cette implication tout au long du cycle de vie d'un produit (dès la conception à sa livraison), permet de voir le PLM comme une plateforme centrique pour capter la « connaissance quotidienne » produite d'une façon collaborative. Utiliser le PLM pour améliorer l'exploitation du savoir-faire de l'entreprise est une idée présentée dans des travaux de recherche. Nous pouvons citer comme exemple [Bissay *et al.* 2008] qui propose une approche composée de 7 étapes :

1. Analyse de l'activité (analyser l'activité de développement des nouveaux produits) ;
2. Identification des connaissances (identifier les connaissances en lien avec le processus de développement des produits) ;

3. Caractérisation des entités techniques (sélectionner les éléments de connaissance pouvant enrichir le méta-modèle de données du PLM) ;
4. Construction de l'espace d'état (construire les cycles de vie des entités techniques de connaissances à partir d'une grille de maturité globale à l'entreprise) ;
5. Identification des rôles/compétences/experts (caractériser les fonctions nécessaires à la mise en œuvre des produits et l'expertise) ;
6. Construction des workflow métiers (définir le processus métiers et les éléments de connaissance qu'il génère) ;
7. Construction des indicateurs (évaluer la performance des éléments de connaissance ainsi que des processus qui les génèrent).

La traçabilité est intégrée dans un PLM pour garder la trace de la connaissance produite au quotidien des collaborateurs « daily knowledge». Par exemple, [Matta *et al.* 2013a] propose d'annoter les changements d'un projet dans un PLM Windchill en exploitant les modifications dans le workflow qui se génèrent après chaque modification ainsi que dans les rapports.
Le « product line systems », ou les lignes de produit, présente aussi un terrain de partage d'expertise et de traçabilité. XTraQue [Jirapanthong & Zisman 2009], à titre d'exemple, est une approche de traçabilité intégrée dans les systèmes de ligne de produit et basée sur des règles pour prendre en charge la génération automatique des relations de traçabilité entre des documents orientés objet et les caractéristiques. Certaines méthodes proposent de récupérer automatiquement les liens de traçabilité des exigences. La plupart

d'entre elles se basent sur la recherche de similarité de représentation entre les exigences et le code source à l'instar de [Turban 2013] et [Tsuchiya *et al.* 2015].

2.5 Synthèse des approches étudiées

Les approches et les méthodes de traçabilité que nous avons discutées permettent de recueillir et de réutiliser d'une manière ou d'une autre l'expérience passée d'un individu ou d'un ensemble d'individus dans une organisation. Nous présentons une vue globale des approches étudiées caractérisées par leur contexte d'utilisation, les techniques de traçabilité proposées, les sources visées ainsi que le support de traçabilité dans le Tableau 2.1 ci-dessous.

TABLE 2.1 – Comparaison des approches de traçabilité.

Principe	Approche	Contexte	Traçabilité Automatique/manuelle	Source de traces	Support de Traçabilité
Structurer la mémoire de l'individu	REX [Malvache & Prieur 1993]	Projet	Manuelle	Structuration des entretiens	Fiche REX
	MEREX [Prax 2012]	Projet	Manuelle	Structuration des entretiens	Fiche MEREX
Structurer la mémoire organisationnelle	DYPKM [Bekhti 2003]	Projet de Conception	Manuelle	Documents de conception	Modèle
	Memory meeting [Matta *et al.* 2013b]	Prise de décision	Semi-automatique	PV des réunions	logiciel
	MEMORAe [Abel *et al.* 2002]	Apprentissage	Semi-automatique	Utilisateur	logiciel

Principe	Approche	Contexte	Traçabilité Automatique/manuelle	Source de traces	Support de Traçabilité
Structurer le comportement de l'utilisateur d'un système d'information	MUSETTE [Champin *et al.* 2004]	Les exigences du projet	Automatique	Fichiers logs	Système de Base de raisonnement
	Traçabilité dans PLM [Bissay *et al.* 2008] [Matta *et al.* 2013a]	Cycle de vie projet	Automatique	Documents du projet	Plugin ou documents
	Traçabilité des exigences [Tsuchiya *et al.* 2015]	Exigences	Automatique	Documents des exigences	Logiciel ou documents
	XTraQue [Jirapanthong & Zisman 2009]	Ligne des produits	Automatique	Documents	Logiciel

Cet état de l'art se veut une étude des connaissances des différentes méthodes et approches impliquées dans les systèmes traçants qui nous mènent à différentes réflexions citées sous forme de points :

— Automatisation de la traçabilité : L'automatisation est un critère de gain en terme de coût et de performance et surtout de disponibilité. Les approches permettant la traçabilité automatique profitent de cet avantage pour une application en permanence de la traçabilité. Cependant, le travail automatique est non supervisé, ce qui peut dégrader la qualité

des résultats obtenus.

— Implication de l'expert dans l'approche de traçabilité : Le travail humain est caractérisé par la précision. Il est garant de qualité mais aussi de subjectivité. Des questions se posent à ce titre : à quel point, dans une entreprise, l'expert (non spécialiste de la gestion des connaissances et de traçabilité) peut être impliqué dans le processus de la traçabilité ? Et surtout, est-il conscient de l'importance de conserver des traces de son travail ? La réponse à ces questions, bien qu'elle diffère d'une organisation à une autre, détermine le degré du succès des approches « manuelles » de traçabilité.

— L'axe central des approches de traçabilité : On observe que la plupart des approches de traçabilité discutées dans la section précédente sont plutôt orientées selon la source de la trace. Autrement dit, la source de la trace (Fichiers log, mémoire de l'expert, documents de conceptions ou d'exigences..) forme l'axe central de l'approche, et toute une approche de traçabilité est construite selon la source. En revanche, dans un processus métier, plusieurs outils sont utilisés. Plusieurs ressources hétérogènes produites sont susceptibles d'être des sources de traces aussi importantes. Les approches de traçabilité qui se basent sur l'analyse du processus métier et dont « l'expert » est leur axe central doivent considérer cette hétérogénéité de supports de traces.

— Traçabilité et personnalisation : Les approches de traçabilité autour des projets ou mémoire de projet ne décrivent pas comment la trace est organisée après l'enregistrement. Si les projets de l'entreprise sont de très longue durée alors les traces à sauvegarder sont encore plus considérables. Si

un expert souhaite chercher une connaissance bien particulière, il devra parcourir tous les enregistrements pour trouver celle qui l'intéresse. Par exemple, l'auteur [Chatzopoulou *et al.* 2011] s'est rendu compte de la problématique des traces générées automatiquement dans le cadre de l'analyse de logs. Il a alors proposé de compléter la traçabilité par un travail de recommandation afin de minimiser le temps de la recherche de l'enregistrement. Par conséquent, si la traçabilité est personnalisée au moment de la capture, il sera par la suite plus facile d'organiser son enregistrement.

2.6 Conclusion

Dans ce chapitre nous avons présenté une analyse des approches de traçabilité dans l'industrie. Les approches proposées sont variées et répondent à des besoins multiples. De même, les sources de traces sont hétérogènes, parfois captées automatiquement ou issues d'entretiens avec des experts. L'apprentissage de l'expérience passée et la construction d'une mémoire collective et partagée est l'objectif de l'ensemble des approches étudiées. Cependant, certaines nécessitent une forte implication de l'individu pour une application réelle de l'approche de traçabilité, ce qui n'est pas le cas dans toutes les organisations. Plus précisément, garder la trace est vu comme une charge de travail en plus pour un expert de l'entreprise. Les axes centraux de certaines approches sont les supports de traces eux-mêmes. Par contre, les supports changent et évoluent, ce qui pose la question de la possibilité pour ces approches de se développer.

Retour à la problématique :

Dans notre problématique nous mettons l'accent sur le fait que le besoin de la traçabilité exige un critère « dynamique » dans le sens

où un système dédié doit réaliser une traçabilité de connaissances d'une façon régulière et « documentaire » vu que nous considérons les documents comme la source la plus importante de la connaissance de l'entreprise. Par rapport à l'objectif que nous souhaitons atteindre nous essayons dans notre approche de réduire le gap entre les points que nous avons observés : l'implication de l'expert, l'automatisation, l'axe central de l'approche et la personnalisation.
Nous proposons donc dans notre approche,**un cadre de traçabilité documentaire, régulier (semi-automatique) autour de l'acteur.** C'est pour cela que la traçabilité doit être personnalisée par acteur considérant la spécificité de son domaine, ses besoins et son profil. **Une traçabilité personnalisée :** une perspective qu'on explore dans nos travaux dans l'objectif d'établir des liens entre la trace et le collaborateur (acteur) qui l'a produite d'une part, et entre la trace et le collaborateur dont il a besoin. Kmiz Dalkir, pour le prendre comme exemple, a proposé de lier la traçabilité aux techniques de profilage [Dalkir 2013] pour permettre d'avoir une traçabilité personnalisée pouvant présenter un support de suivi des interactions d'un individu.
Nous présentons figure 4.1 ci-dessous les différentes dimensions que nous considérons nécessaires pour que l'approche que nous proposons soit efficace et pertinente.

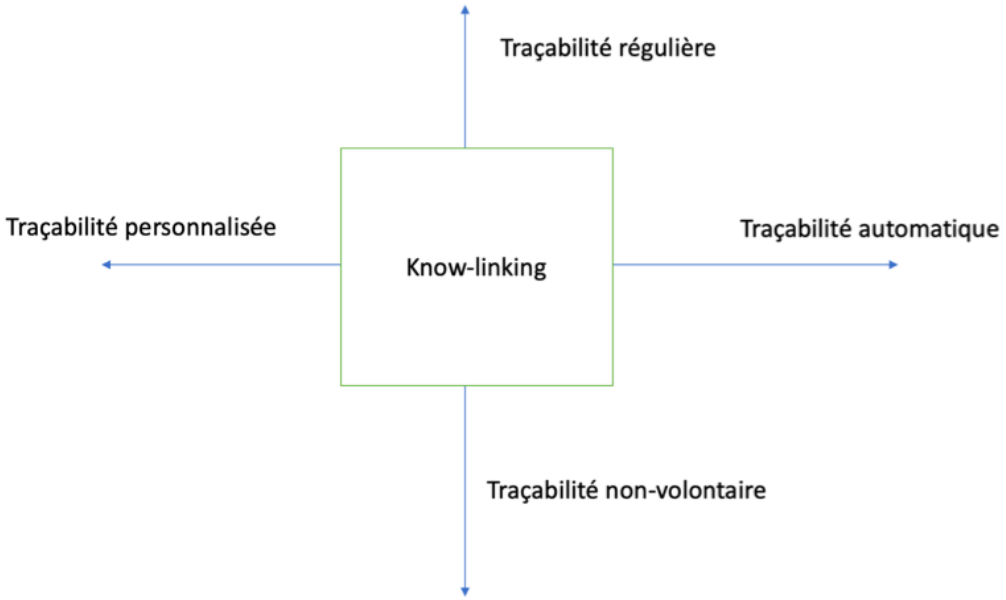


FIGURE 2.2 – Les dimensions du Know-linking

Chapitre 3

Profilage

«**Le profil adéquat peut être vu de face**» Gaeten Faucer

Sommaire

3.1 Introduction

À l'issue de nos travaux de recherche en Traçabilité comme support pour un partage de connaissance, nous nous sommes posé la question de la possibilité d'ajouter une dimension de personnalisation dans notre approche. Nous avons trouvé la réponse dans le champ de profilage qui permet d'effectuer un partage ciblé de la connaissance. Une étape importante pour résoudre les problèmes du partage de la connaissance est d'analyser et de comprendre le besoin des collaborateurs ainsi que les périmètres de leurs contributions à la production de la connaissance. Cela va permettre de repérer à la fois leurs besoins en connaissances ainsi que celles qu'ils produisent.

Dans le cadre de l'entreprise, plusieurs collaborateurs partagent les mêmes missions ou les mêmes tâches au quotidien, ce qui explique qu'ils partagent le même besoin en connaissances. Ces sous-groupes ayant des caractéristiques communes dans une entreprise à large échelle peuvent être regroupés selon ces caractéristiques dans des structures génériques appelées «profils».

Bien que la création des profils ou «profilage» de personnes soit largement utilisée dans les domaines en rapport direct avec la clientèle tels que le marketing, le secteur bancaire ou la finance, etc., l'histoire du profilage est plus ancienne que «l'ère du data».

Dans ce chapitre nous étudions les différentes définitions des termes «profil» et «profilage». Nous discutons également de quelle manière les techniques du profilage existantes dans la littérature scientifique peuvent être adoptées dans notre méthodologie de partage de connaissance afin d'assurer un axe de personnalisation de notre approche «Know-linking» centrée sur les collaborateurs.

3.2 Profil : définition et caractéristiques

Dans le dictionnaire Larousse, un profil est défini comme un «ensemble de caractéristiques qui définissent fondamentalement un type de chose, configuration de quelque chose à un moment donné». Dans l'ère du «data», ces caractéristiques peuvent apparaître sous forme de données relatives à un sujet. Dans ce cas, un profil peut être défini comme «un ensemble de données corrélées qui représentent un sujet «humain ou non humain, individu ou groupe» [Hildebrandt & Gutwirth 2008]. Bien que la première définition linguistique repose sur le mot «caractéristique», la seconde met l'accent sur la représentation.
Un profil peut concerner un groupe ou un individu [Hildebrandt & Gutwirth 2008].

— Un profil de groupe : est une représentation d'un ensemble d'entités ayant les mêmes «attributs» [Ferraris *et al.* 2013], donc tous les membres du groupe partagent les mêmes caractéristiques [Ferraris *et al.* 2013]. Par exemple, tous les clients d'une banque qui sont encore étudiants partagent les mêmes caractéristiques : statuts, tranche d'âge, même situation financière, etc. Par conséquent, un profil «étudiant» dans une banque englobe toutes les personnes qui partagent ces caractéristiques.
— Un profil de groupe non distributif : les membres d'un groupe peuvent ne pas partager les mêmes caractéristiques. Un profil de groupe non distributif [Ferraris *et al.* 2013] caractérise les membres d'un groupe qui ont seulement quelques caractéristiques en commun et non pas toutes. Par exemple des personnes ayant un risque élevé d'attraper le covid-19 peuvent être profiler sur l'occurrence d'un certain nombre de facteurs comme : vacciné ou non, se retrouver dans un endroit fermé

avec plein de monde, être en contact récent avec un cas positif...etc. En revanche, un membre de ce groupe peut avoir un seul attribut en commun.

— Un profil personnalisé : une personne peut être identifiée dans un groupe par son profil qui représente l'ensemble d'attributs appartenant à cette personne. Ce profil «personnalisé» ou «individuel» [Ferraris *et al.* 2013] peut être utilisé pour identifier un individu parmi un groupe ou pour déduire certaines de ses caractéristiques (Figure 3.1).

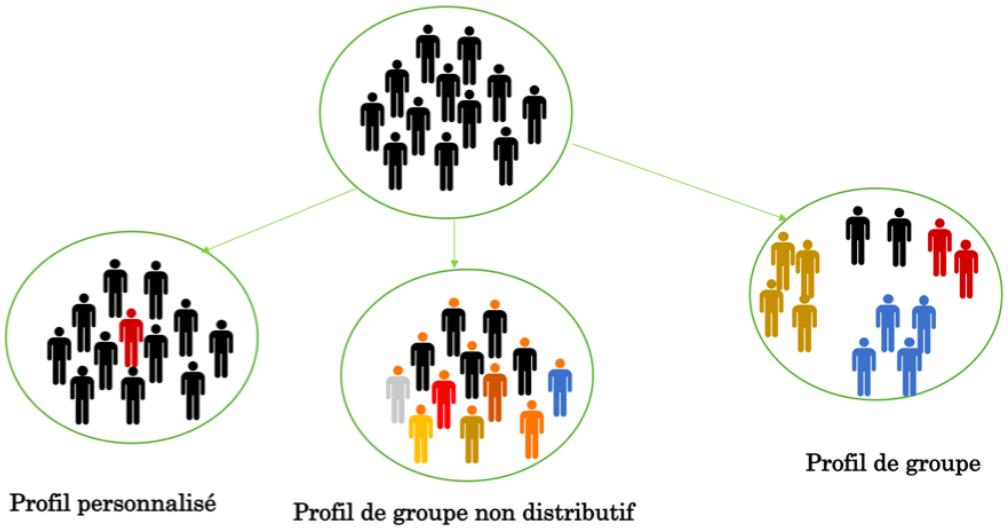


FIGURE 3.1 – Types des profils

D'autres travaux comme ceux de [Lashkari *et al.* 2019] soulignent que les profils d'utilisateur représentent un ensemble de patterns qui peuvent être communs à plusieurs utilisateurs. Dans la même étude l'auteur pense que la connaissance acquise par un utilisateur donne une «forte indication sur ses pensées et permet de prédire ses intentions». Un exemple donné par [Ortigosa *et al.* 2014] avance l'idée que le profilage des interactions au sein des réseaux sociaux comme Facebook et Twitter permet de prédire la personnalité de l'utilisateur.

En analysant toutes les définitions étudiées, un profil est donc une représentation des caractéristiques d'un individu dit «unique» ou d'un ensemble de caractéristiques communes entre plusieurs individus.

3.3 Profilage : définition

L'une des définitions les plus anciennes du profilage dans la littérature scientifique, et présentée par [Marx & Reichman 1984] en 1984, souligne un rapport d'opposition entre «le profilage» et «le matching», et dévoile une logique sous-tend «une logique inductive dans la recherche d'indices qui augmenteront la probabilité de découvrir des infractions par rapport aux recherches aléatoires» [Marx & Reichman 1984]. Le profilage, d'après la même définition, permet de «corréler un certain nombre d'éléments de données distincts afin d'évaluer à quel point une personne ou un événement se rapproche d'une caractérisation ou d'un modèle d'infraction prédéterminé» [Marx & Reichman 1984]. En analysant cette définition nous pouvons voir qu'elle est plus liée aux domaines juridiques (d'après le background de l'auteur). En 1993, Roger Clarke [Clarke 1993] a introduit le profilage comme une «technique de surveillance des données» ou «dataveillance technique», ainsi qu'«un processus de création et d'utilisation d'un profil» [Clarke 1993]. Après l'évolution importante des techniques de prédiction qui sont basées sur la découverte des «patterns» ou des «modèles», le profilage est vu comme «la découverte de modèles qui représentent des connaissances permettant d'anticiper des événements futurs en fonction de ce qui s'est passé» [Ferraris *et al.* 2013], [Hildebrandt & Gutwirth 2008]. En 2009, Mireille Hildebrandt a complété ses anciens travaux en définissant le profilage comme «profiling the European Citizen» [Hildebrandt &

Gutwirth 2008] pour mettre en valeur l'importance de la prédiction permise par le profilage. Elle présente de nouveau le profilage comme la «découverte de modèles qui présentent des connaissances et qui permettent d'anticiper des événements futurs sur la base du passé» [Hildebrandt 2009].
Par ailleurs, il existe des définitions officielles concernant le profilage ou «profiling». Le CoE (le conseil d'Europe), par exemple, définit le profilage comme étant une «technique de traitement automatisé de données qui consiste à appliquer un «profil» à un individu, notamment pour prendre des décisions le concernant ou pour analyser ou prédire ses préférences, comportements et attitudes personnels» (paragraph 1, CM/Rec (2010)13) [CoE]. Une autre définition officielle est donnée par le RGPD (Règlement général sur la protection des données), dans laquelle le profilage est défini comme «un traitement automatisé destiné à évaluer certains aspects personnels relatifs à cette personne physique ou analyser ou prédire notamment les performances de la personne physique au travail, sa situation économique, sa localisation, sa santé, ses préférences personnelles, sa fiabilité ou son comportement» (article 20, Proposition de RGPD, 2012). Le mot profilage a été employé dans plusieurs contextes comme la biologie ou la criminologie, mais aussi dans la documentation et la clientèle, etc. Nous consacrons les pages suivantes à leur présentation.

3.3.1 Profilage : une technique inspirée d'un phénomène naturel

L'histoire du profilage est plus ancienne que les sciences d'analyse de la clientèle ou la criminologie, des domaines d'application de techniques de profilage moderne. En 1999, Van Brakel a, par

exemple, présenté la relation entre la biologie et la théorie de l'information comme deux domaines développés en un domaine intégré faisant partie des sciences de la vie [Hildebrandt & Gutwirth 2008]. Si nous revenons à l'origine du terme nous constatons que les études biologiques ont démontré que le profilage est un phénomène naturel. En effet, pour garantir leur survie, les organismes non-humains doivent s'adapter à leur environnement. Ils essaient donc en permanence de collecter des informations sur l'environnement qui les entoure à travers les différentes interactions qui ont lieu avec ce dernier. Cette phase de collecte et de traitement de l'information est caractérisée par «l'inconscience» de l'organisme non-humain, donnant naissance au terme «profilage organique» [Hildebrandt & Gutwirth 2008] [Ferraris *et al.* 2013]. En ce qui concerne l'être humain, ce processus de profilage est différent car ce dernier est caractérisé par sa réflexion et son intention. Ainsi, le profilage humain se fait automatiquement et dans une large mesure. Autrement dit, l'être humain s'inscrit dans un processus d'apprentissage ce qui transforme les informations collectées et analysées par des habitudes préalablement définies par la conscience humaine [Hildebrandt & Gutwirth 2008]. Par conséquent, ce profilage «automatique» est le résultat d'un processus d'apprentissage [Ferraris *et al.* 2013]. En d'autres termes, les humains agissent inconsciemment ou involontairement parce que ces actes sont des habitudes préalablement définies par la conscience humaine [Ferraris *et al.* 2013]. Autre que le profilage organique et humain il existe une troisième famille : le profilage machine. Ce type de profilage n'implique aucune intention ou conscience, ce qui le rapproche du profilage organique [Hildebrandt & Gutwirth 2008] [Ferraris *et al.* 2013]. Les trois familles présentées permettent d'introduire une catégorisation du profilage : profilage non

automatisé, autrement dit un profilage manuel assuré par un individu, automatisé s'il est assuré par un algorithme et autonome s'il est effectué sans implication de l'être humain.

3.3.2 Profilage : un courant de la criminologie

Le terme profilage est souvent lié à la science de la criminologie. Il est d'ailleurs adopté par des structures comme le FBI aux Etats-Unis car il représente en effet le troisième courant de la science de l'investigation [Bond]. Dans la criminologie, le profilage est défini comme «le processus d'identification des traits de personnalité, des tendances comportementales et des variables démographiques d'un délinquant en fonction des caractéristiques du crime» [Kocsis 2006], une définition qui met l'accent sur les aspects psychologiques et comportementaux de l'être humain.
[Holmes & Holmes 2008] ont déterminé essentiellement trois objectifs principaux du profilage criminel :

- Fournir aux autorités d'application de la loi une évaluation socio-psychologique du criminel ;
- Donner à la police une «évaluation psychologique des effets personnels trouvés en possession du délinquant» ;
- Fournir des conseils et des stratégies pour les interrogatoires.

Dans un cadre plus globalisé, le profilage criminel est vu comme l'ensemble des techniques psychologiques appliquées afin d'identifier l'auteur d'un crime. Ces techniques psychologiques permettent d'établir un lien entre le passé du criminel et la nature du crime. En d'autres termes, l'analyse du passé d'un criminel détermine les aspects de la criminalité et aide à analyser les preuves de la scène du crime.

3.3.3 Profilage des documents

La recherche d'un document est devenue une tâche de plus en plus compliquée, surtout quand il s'agit d'une recherche sur internet. Le nombre de documents publiés au quotidien, les sujets qui sont très variés ainsi que d'autres facteurs ont poussé les chercheurs à appliquer le profilage pour faciliter l'accès aux documents et pour trouver un format de documents à rechercher sur la base de leur contenu. Des études comme [Sauban & Pfahringer 2003] se base sur le modèle de Lee [Lee 2001], un modèle psychologique qui prend en compte trois aspects différents : «les gens sont capables non seulement d'affirmer qu'un document donné porte sur un sujet donné, mais aussi qu'un document ne traite pas ce sujet», «les humains sont capables de prendre des décisions non compensatoires», autrement dit ils peuvent décider si un document porte sur un sujet ou non sans nécessairement avoir à lire tout le document. «Les personnes sont capables de donner une réponse avec un niveau de confiance», ainsi ils peuvent affirmer qu'un document traite soit définitivement d'un sujet, soit qu'il est simplement lié à un sujet. En se basant sur ces aspects, les mots d'un document sont évalués au fur et à mesure par une fonction de probabilité pour savoir s'ils appartiennent ou non à un sujet donné. Les auteurs définissent des profils de documents qui sont entre autres la somme partielle des logs des cotes «log-odds» des mots d'un document. Si les cotes d'un document par rapport à une catégorie bien déterminée sont suivies, au fur et à mesure les mots seront transmis au système. Une fois les profils préparés, les auteurs procèdent à une classification automatique des documents selon ces profils.

Un autre travail de construction de profil de document a été mené

par Antonio Guillén et al [Guillén *et al.* 2017]. Ils utilisent les techniques de traitement de langage naturel (Natural language Processing), et de l'extraction de l'information (à base de pattern), afin d'analyser le contenu d'un document et de générer automatiquement un profil de document. D'après la même approche, le profil de document est l'ensemble de ses propriétés tels que : son type, une estimation de l'âge, son idéologie, sa langue, la liste des sujets, sa région, son résumé, ses mots clés. Selon la base de ses propriétés le document sera classé [Guillén *et al.* 2017].

Ces approches de profilage ont prouvé leur efficacité pour faciliter l'accès aux documents, tout comme l'indexation sémantique qui considère le contenu du document. Cependant, pour une démarche de résolution d'un problème de partage de documents entre plusieurs utilisateurs, ces approches présentent des limites. Notons par exemple la facilité pour un employé d'identifier les thématiques en rapport direct avec son métier. Il existe d'autres thématiques qui sont en rapport avec celles de base et qui sont non-considérées par ces approches. Autrement dit, un document du bilan d'achat des pièces mécaniques sera classé sous la thématique mécanique parmi plusieurs autres documents. Si le responsable d'achat de l'entreprise est celui qui demande ce genre de documents, doit-il fouiller tous les documents de la mécanique pour retrouver ce qui l'intéresse ? Ce n'est qu'un exemple parmi tant d'autres qui prouve qu'il faut aller plus loin que créer des profils de documents et les classer selon des thématiques, à la recherche de liens entre les thématiques qui impactent le partage des documents. Pour pouvoir identifier les thématiques ainsi que les rapports qui existent entre eux, il faut avoir une représentation structurelle qui formalise ces liens.

Le profil métier peut structurer ces ensembles appelés profil de collaborateur. Les collaborateurs sont les utilisateurs des systèmes

d'information de l'entreprise. Ils peuvent être repérés par des «profils d'utilisateur».

3.3.4 Profilage des utilisateurs (user profiling)

En plus de la justice et de la criminologie, le profilage a été adopté par d'autres secteurs comme le marketing, l'assurance, et le secteur bancaire, des secteurs où la satisfaction de la clientèle est la clé de la réussite. Dans ce contexte, les clients sont en général considérés comme des utilisateurs de services ou de plateformes. Leurs données sont collectées et analysées dans le but de comprendre leurs besoins et de trouver un moyen de les satisfaire. Le profilage des utilisateurs est un processus d'identification de données en lien avec les intérêts des utilisateurs [Eke *et al.* 2019]. Ces données peuvent servir à la personnalisation des services ou à la proposition de recommandations.
Profiler les utilisateurs c'est aussi créer des profils d'utilisateurs qui sont «des représentations virtuelles» de chaque utilisateur incluant une variété d'informations le concernant comme : des informations personnelles, leurs intérêts ainsi que des données à propos de leurs préférences [Eke *et al.* 2019]. Yang [Yang 2010], dans ses travaux de recherche, a souligné que le profil de l'utilisateur aide à «résumer une grande quantité d'informations» sur l'utilisateur dans l'objectif de personnaliser la recherche de l'information et la recommandation des produits.
D'une façon générale, le profilage des utilisateurs a pour objectif de collecter des informations sur les centres d'intérêt d'un utilisateur durant un laps de temps bien déterminé dans le but d'améliorer la qualité d'accès aux informations et d'identifier les centres d'intérêt des utilisateurs [Eke *et al.* 2019].
Les travaux menés par Christopher Ifeanyi Eke et al [Eke *et al.* 2019] d'études et d'enquêtes sur les travaux existants de profilage dans

leurs survey publié en 2019, les ont conduit à modéliser le «user profiling» sous la forme d'une taxonomie décomposée en plusieurs niveaux. Cette taxonomie présentée dans la figure 3.2 ci-dessous récapitule d'une façon hiérarchique et organisée les différents travaux de profilage d'utilisateurs qui existent.

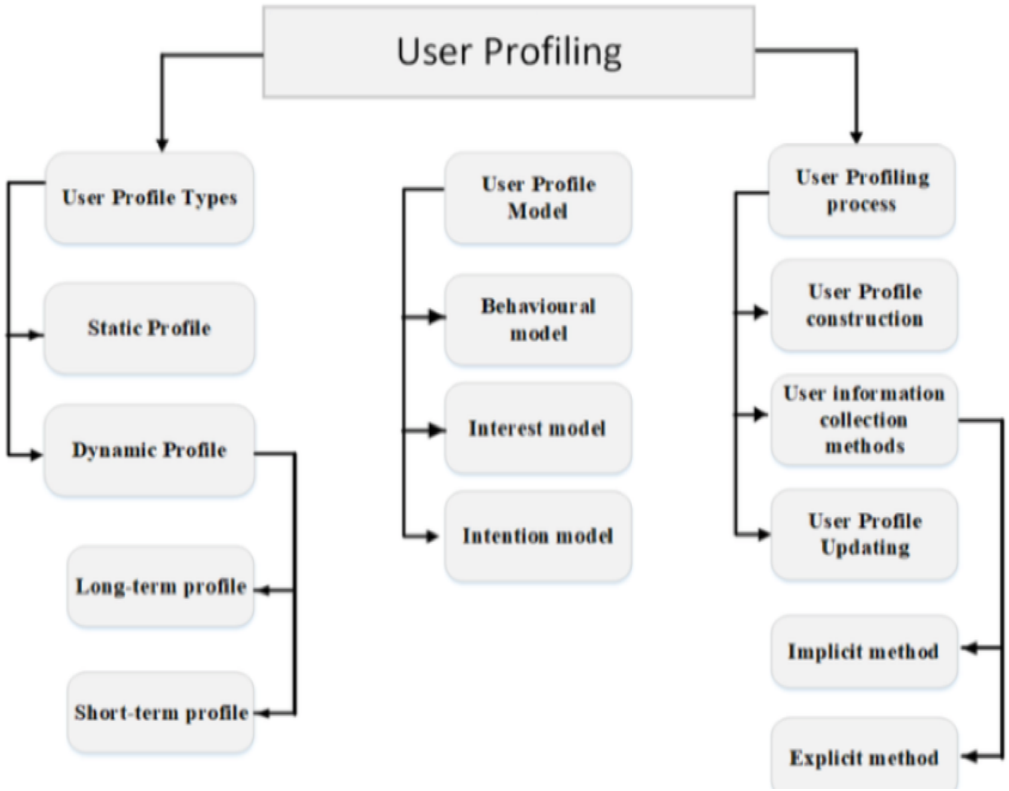


FIGURE 3.2 – Taxonomie profilage des utilisateurs [Eke *et al.* 2019]

3.4 Récapitulatif des définitions présentées

Le profilage est souvent utilisé dans plusieurs domaines d'application tels que l'analyse des événements, la recommandation des services ou l'inférence des attributs. L'ensemble des définitions présentées du terme profilage, et malgré les différences au niveau de l'application, se résume en certains points :

— La recherche des caractéristiques : le profilage permet de construire des sous-ensembles de représentations d'individus partageant les mêmes caractéristiques ;
— Représentation des informations : la construction des profils est la construction d'une représentation d'un ensemble d'informations ;
— La finalité du profilage : le regroupement des individus dans des sous-groupes sur base de leurs caractéristiques communes permet d'analyser et de comprendre leur comportement et ainsi de l'anticiper et d'en prédire de nouveaux. Par ailleurs, la personnalisation d'un service ou la recommandation sont possibles suite à la compréhension éventuelle de la trajectoire des sous-ensembles à travers le profilage ;
— Un lien étroit avec la personnalisation : bien que la personnalisation fasse partie des finalités du profilage, le lien entre eux est plus profond dans la mesure où on utilise le terme «personnalisation» pour désigner «le profilage» dans certains travaux de recherche.

3.5 Processus de profilage

L'application du profilage est la mise en place d'un processus composé essentiellement de trois phases :

1. La création du profil [Eke *et al.* 2019] : travail de collecte d'informations sur l'individu à travers une interaction directe qui peut être manuelle ou automatique.
2. La collecte d'informations [Eke *et al.* 2019] : collecte d'informations concernant un individu en particulier. Ces informations peuvent être collectées manuellement (informations

introduites directement par l'utilisateur), ou automatiquement en se basant sur des agents «intelligents» comme les cookies, les «agents logiciels»..etc.

3. La mise à jour du profil [Eke *et al.* 2019] : la dernière phase du processus de profilage est la mise à jour des profils. À l'issue de la réussite de la création du profil, la mise à jour consiste à enrichir les informations du profil par des requêtes ciblées aux systèmes [Eke *et al.* 2019]. Des requêtes basées sur les intérêts du profil et les mots clés définis aux préalables.

3.6 Moyens du profilage

Différents moyens de profilage existent dans la littérature scientifique et peuvent être classés dans plusieurs catégories selon la nature de la collecte de l'information utilisée. Le profilage peut se faire d'une façon implicite ou explicite, automatique ou non. Dans la partie suivante nous détaillons chaque catégorie de profilage.

3.6.1 Profilage implicite/explicite

Les informations peuvent être capturées de deux façons différentes afin de générer les profils, d'une manière implicite ou explicite. Si la collecte d'informations se fait par le biais d'une intervention «directe» et «manuelle» de l'expert pour capturer les informations correspondantes, ce profilage est appelé «explicite» [Eke *et al.* 2019]. Voici des exemples de types d'informations pouvant être capturées pour ce profilage : le nom d'un utilisateur, son adresse, son statut social, son numéro de téléphone ou bien sa date de naissance, etc. Ces informations peuvent être capturées à l'aide de techniques comme les questionnaires, les «ratings» ou bien des formulaires de «feedback» [**?**, **?**]. Danny Poo [Poo *et al.* 2003] a souligné que les informations et les caractéristiques «statiques»

collectées implicitement conduisent à un profilage statique, autrement dit la construction de profils statiques qui ne changent pas au fil du temps, ce qui dégrade la qualité du profilage [Kanoje *et al.* 2015].

D'autre part, si les informations sont capturées d'une façon implicite, ce qui exclut l'intervention «directe» de l'utilisateur, ces informations sont capturées inconsciemment. Dans ce cas le profilage est dit «implicite». Ce travail de collecte d'informations implicite présente un point faible sur le plan éthique si les utilisateurs ne veulent pas dévoiler certaines informations qui les concernent. Danny Poo [Poo *et al.* 2003] considère que ce profilage est dynamique, ainsi il permet de construire des profils «dynamiques». Les informations collectées concernent plus le comportement d'un utilisateur, ce qui explique l'appellation de ce profilage dans certains travaux comme «behavioral profiling», «Adaptive Profiling» ou encore «Ontological Profiling» [Kanoje *et al.* 2015]. Ces informations implicites peuvent être regroupées à travers des techniques de fouille de données (Datamining) ou à travers des agents «intelligents». Ce type de profilage est souvent lié aux techniques d'apprentissage automatique (Machine Learning). D'autres techniques peuvent également être appliquées telle que : «Rule based filtering», où le système se base sur les informations d'un utilisateur afin de construire des règles sous la forme de «if this then that», ou bien le «Collaborative filtering» qui se base sur l'analyse du passé d'un utilisateur afin de déterminer son appartenance à un sous-groupe d'utilisateurs ou «Content based filtering techniques», où un flux d'informations est comparé aux informations d'un profil afin de détecter ce qui peut l'intéresser [Kanoje *et al.* 2015].

3.6.2 Profilage manuel/automatique

La création des profils peut se faire manuellement par un individu qui peut être un expert ou même par les utilisateurs qui introduisent leurs informations directement et définissent leurs centres d'intérêt, i.e des méthodes de profilages qui utilisent les réseaux sociaux tels que YouTube ou LinkedIn [Grčar *et al.* 2005] [Eke *et al.* 2019]. L'utilisateur peut ne pas fournir toutes ses informations pour garder une part de confidentialité, ce qui peut générer des profils «incomplets». Cependant, les méthodes de profilage automatique se basent sur des algorithmes pour collecter et générer des profils dynamiques et «comportementaux. Les mécanismes de profilage automatique peuvent être classés en trois paradigmes : «statistical keyword analysis», «social filtering algorithms» et «machine learning techniques» [Soltysiak & Crabtree 1998].

— Statistical keyword analysis, qui se repose sur des techniques de recherche d'informations classiques. Avec cette méthode, les mots-clés sont analysés isolément, le contexte de l'information n'est pas pris en considération ce qui affecte la qualité des profils [Schiaffino & Amandi 2000].

— Les algorithmes de filtrage social ont généralement besoin d'une large communauté d'utilisateurs pour fonctionner efficacement [Schiaffino & Amandi 2000].

— La troisième méthode est basée sur les algorithmes d'apprentissage automatique pour générer des profils d'utilisateurs [Van Otterlo 2013], [Schiaffino & Amandi 2000]. Les résultats du profilage automatique sont plus efficaces et plus pertinents [Eke *et al.* 2019].

3.6.3 Les méthodes de profilage hybride

Le terme hybride peut être utilisé pour désigner les méthodes de profilage combinant deux techniques différentes, par exemple implicite et explicite ou automatique et manuelle. Ces techniques hybrides profitent des avantages de l'union de deux techniques pour de meilleurs résultats.

3.6.4 Comparaison des méthodes de profilage

Les méthodes de profilage peuvent être comparées selon les critères que nous avons détaillés dans la section précédente et que nous avons illustrés dans le tableau comparatif 3.1 ci-dessous.

TABLE 3.1 – Tableau comparatif des méthodes de profilage.

Approche	Implicite	Explicite	Automatique	Manuelle
Statistical keyword analysis [Schiaffino & Amandi 2000]		X	X	
Social filtering algorithms [Schiaffino & Amandi 2000], [Ismail *et al.* 2018]	X		X	
Machine learning techniques [Van Otterlo 2013]	X		X	
Rule based filtering [Kuflik *et al.* 2003]		X	X	
Collaborative filtering [Stefanidis *et al.* 2018]	X		X	
Content based filtering [Geetha *et al.* 2018]	X		X	
Questionnaires [Klement 2015]		X		X

3.7 Algorithmes de profilage

Les algorithmes de profilage permettent en premier lieu de collecter les informations puis de générer des profils sous forme d'une représentation ou d'un modèle. Dans la littérature scientifique, plusieurs algorithmes sont adoptés pour pouvoir générer et représenter les profils. Selon la représentation du profil, ces algorithmes sont catégorisés. À titre d'exemple, une classification selon le mécanisme de représentation des profils présenté en 2005 par [Godoy & Amandi 2005] classifie les algorithmes en : représentation des documents et algorithme d'apprentissage supervisé et non-supervisé. En 2019, une autre étude [Eke *et al.* 2019] a classifié les algorithmes de profilage en cinq familles : «les algorithmes de voisinage», «les algorithmes de machine learning», «les algorithmes basés sur une ontologie», «le filtrage» et «les modèles statistiques». Après avoir étudié les algorithmes présentés dans ces travaux, nous nous basons sur la catégorisation des profils en détaillant dans la partie suivante les différents algorithmes de profiling ainsi que leurs principes :

3.7.1 Neighbourhood based

L'idée de cet algorithme est basée sur l'hypothèse suivante : «il existe des amis partageant les mêmes centres d'intérêt», alors ce groupe est appelé «neighbourhoud» [Eke *et al.* 2019]. Les modèles basés sur le «neighbourhood» calculent généralement la similitude entre les utilisateurs ou les éléments et utilisent ces similitudes pour prédire des évaluations inconnues. Ainsi, le comportement d'un utilisateur peut être prédit selon le comportement de ses «neighbourhood». Il faut savoir que le processus de création des «neighbourhood» est un processus de création d'un modèle qui conduit à la recommandation collaborative [Eke *et al.* 2019].

L'objectif principal du «Neighbourhood» est donc de déterminer pour chaque utilisateur I une liste ordonnée d'utilisateurs J : Mb = (M1,M2,...Mj), tel que b Mb où (b, M1) = maximum et Sim(b,M2) = prochain max, etc. À l'aide d'une fonction de distribution cumulative, [Jurgens *et al.* 2015], a prouvé qu'il est possible de prédire la localisation individuelle à partir du «neighbourhood», notamment le plus proche voisin. En se basant sur des informations sur Twitter, l'auteur a indiqué que la moitié des individus ont des voisins qui ont révélé leur emplacement à proximité. Par conséquent le modèle à réussi à identifier la localisation de la maison de l'utilisateur (d'après les résultats de l'expérimentation).

3.7.2 Machine Learning

Le machine learning est une branche de l'intelligence artificielle. Arthur Samuel l'a définie comme un «domaine d'étude» qui offre aux ordinateurs la capacité d'apprendre [Mahesh 2020].
Le machine learning permet d'avoir une interprétation des informations extraites à partir d'une large quantité de données. L'objectif du machine learning est ainsi d'apprendre automatiquement à partir d'un volume de données [Mahesh 2020]. Cet apprentissage peut se faire de deux manières différentes : supervisé ou non supervisé. L'apprentissage supervisé est un exercice d'apprentissage automatique qui consiste à mapper une entrée à une sortie en se basant sur des exemples de pairs entrée-sortie [Mahesh 2020]. Ces opérations de «mapping» sont généralement des opérations de classifications que le système a appris à réaliser à partir d'un ensemble de données entrées au préalable. Ces données sont appelées «training» ou données d'apprentissage [Eke *et al.* 2019] [Kotsiantis *et al.* 2007].

Contrairement à l'apprentissage supervisé, les approches du machine learning non supervisé ne se basent pas sur les données d'apprentissage. Les algorithmes explorent des données pour découvrir des modèles les agrégeant [Mahesh 2020]. Pour reconnaître les classes des nouvelles données, les algorithmes de machine learning non supervisés se basent sur des caractéristiques extraites à partir de données déjà explorées [Mahesh 2020]. De cette façon l'algorithme apprend des caractéristiques à partir de données entrées au fur et à mesure. Il les applique si une nouvelle donnée est introduite et ainsi de suite. Les algorithmes de l'apprentissage supervisé ou non supervisé permettent de découvrir d'une façon dynamique les profils à partir d'une entrée de données. Les techniques de profilage à base des algorithmes de machine learning supervisé comme K-Nearest Neighbour, Naive Bayes et Support Vector Machine, sont entrainés par un jeu de données au préalable. Le K-Nearest neighbour est souvent utilisé pour des problèmes de classification ou de régression. K-Nearst Neighbour ou K-plus proche voisin se base sur une mesure de similarité. Il suffit donc de définir ce que veut dire «semblable» dans le contexte des entrées et de définir aussi «l'influence» de ces voisins sur la prédiction de la cible pour une entrée de test [Delalleau & Larochelle 2007]. L'objectif de cet algorithme est de prédire pour une entrée x la cible correspondante. En effet, il détermine les k- plus proches voisins de x et ce selon une métrique (calcul d'une distance euclidienne, norme L2, ou de façon plus générale la norme Lp de Minkowski).
La prédiction en cas de classification correspond à la classe majoritaire parmi les k plus proches voisins [Delalleau & Larochelle 2007]. Ainsi l'algorithme support vector machine permet de répondre aux enjeux de la classification et de la régression, étant inspiré de la théorie statistique de l'apprentissage de Vladimir [Vapnik 1999].

Les SVM reposent sur l'existence d'un classificateur linéaire (hyperplan) dans un espace approprié [Mohamadally & Fomani 2006]. Le principe de l'algorithme est d'arriver à tracer un séparateur optimal des données et de maximiser la distance entre ces deux classes. Les points les plus proches à l'hyperplan tracent deux vecteurs de supports. L'hyper plan optimal recherché est celui qui maximise la marge entre les deux vecteurs de support (figure 3.3) . L'approche Naive bayes [Webb *et al.* 2010] [Osisanwo

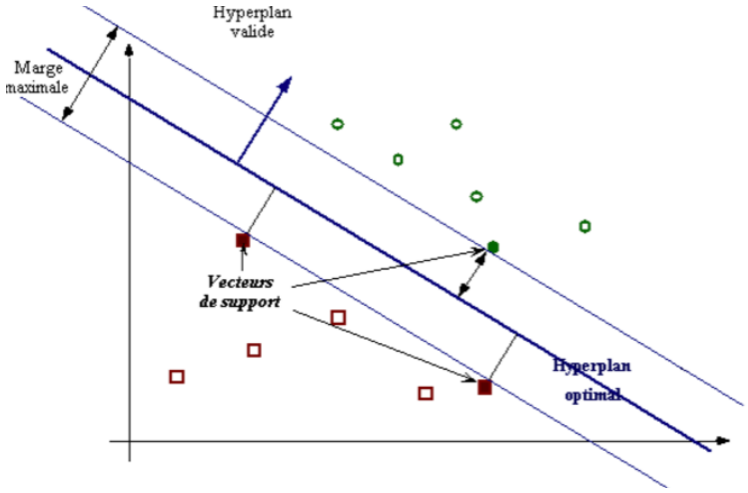


FIGURE 3.3 – Support Vector Machine [Mohamadally & Fomani 2006]

et al. 2017] [Salperwyck & Lemaire 2011] est également exploitée dans les algorithmes de machine learning qui permet de réaliser un exercice de classification supervisée à base d'un modèle probabiliste. Le «classifieur» suppose que les prédicateurs sont indépendants conditionnellement de la variable cible et que la présence d'une caractéristique n'est pas liée à la présence d'une autre. L'algorithme utilise alors la formule de Bayes :

$$P(C|X) = \frac{P(C)P(X|C)}{P(X)}$$

FIGURE 3.4 – Naïve Bayes

L'algorithme Naïve bayes a prouvé son efficacité dans plusieurs

champs d'application tels que : la classification des textes, l'analyse médicale et les systèmes de gestion de performance [Ghosh *et al.* 2020], [Domingos & Pazzani 2004] [Rish *et al.* 2001]. Cependant, les algorithmes de machine learning non-supervisés tel que K-means sont dotés de plus d'autonomie dans le sens où ils ne sont pas entrainés et qu'ils apprennent à découvrir les modèles sans les données d'apprentissage. À titre d'exemple, l'agorithme K-means répond aux enjeux du «clustering», l'idée de l'algorithme étant de créer un nombre K de clusters. Au départ, un nombre de centres k doit être choisi. L'étape suivante consiste à associer chaque point représentant une donnée au centre le plus proche parmi les k-centres de départ. À l'issue de l'affectation des points, les regroupements ou «clusters» devront être crées. À ce stade nous avons besoin de re-calculer k nouveaux centres comme Bary centre des «clusters» résultant de la précédente [Mahesh 2020].
Les algorithmes du machine learning, supervisés ou non, permettent donc de créer les profils d'utilisateurs en partant sur des données labélisées ou dites d'apprentissage pour les algorithmes supervisés, ou «from scratch» pour les algorithmes non supervisés. L'exploration et le clustering sont maintenus jusqu'au bout en se basant sur le retour (feedback) de l'utilisateur. Ces algorithmes, dont nous avons détaillé le principe, participent à l'automatisation de la génération des profils, ce qui a encouragé plusieurs travaux de recherche à les adopter. Par exemple, une étude récente menée par [Ying *et al.* 2018] montre leur exploitation pour la vérification de la paternité «authorship» et la détection d'un compte compromis. Les auteurs ont utilisé la classification de K-NN comme algorithme d'apprentissage qui aide à la mise à jour dynamique des profils. La précision du modèle ainsi que les attributs des profils sont améliorés par une mise à jour régulière du «classifier» d'apprentissage.

3.7.3 Basé sur l'ontologie

Gideon Harvey [Harvey 1973] définit l'ontologie en philosophie comme une science ou étude «métaphysique» de l'être. Formellement, une ontologie est tuplet O = C,T,Rc,Rt où C est une liste de concepts, T un ensemble de termes, Rc désigne les relations entre les concepts et Rt désigne l'ensemble des relations entre les termes tel que Rc : C × C est la relation d'ordre partiel sur C définissant la hiérarchie entre les concepts, Rc (c1, c2) signifie que c1 est plus général que c2, et Rt : C T est une fonction d'association d'un terme préféré à un concept [Di Jorio *et al.* 2007].
[Guarino & Giaretta 1995], dans son travail sur la représentation de la connaissance, s'est basé sur l'expression de Tom Gruber pour décrire le modèle ontologique, l'expression qui dit «une spécification explicite d'une conceptualisation». En ingénierie de connaissances, l'ontologie est souvent utilisée pour représenter les connaissances. [Roche 2005] a donné un résumé des définitions de l'ontologie en ingénierie de connaissances comme «ontologie définie pour un objectif donné et un domaine particulier». Une ontologie est pour l'ingénierie des connaissances une représentation d'une modélisation d'un domaine partagé par une communauté d'acteurs. «Objet informatique défini à l'aide d'un formalisme de représentation, elle se compose principalement d'un ensemble de concepts définis en compréhension, de relations et de propriétés logiques» [Roche 2005]. La représentation ontologique est la base de certaines techniques de profilage où les profils et leurs centres d'intérêt sont représentés sous forme d'une ontologie [Eke *et al.* 2019]. Par exemple l'ontologie OUPA [Han *et al.* 2013] qui se construit automatiquement pour maintenir la représentation des intérêts personnels des profils des utilisateurs. Dans le même contexte, les technologies du web sémantique permettent de construire des profils d'utilisateurs

enrichis en sémantique en se basant sur les pages web balisées par des méta-données sémantiques [Berners-Lee *et al.* 2001].
Le travail de Grimnes [Grimnes 2003] met l'accent sur l'impact du web sémantique sur la personnalisation et propose de fournir des informations structurées, ce qui permet de réduire l'ambiguïté et de donner des références utiles aux informations de base sous la forme d'ontologies. «Le Web sémantique pourrait aider à résoudre les problèmes fondamentaux qui rendent l'apprentissage automatique sur le Web difficile à appliquer et par conséquent surperformer l'apprentissage à partir de documents non structurés» [Grimnes 2003] [Godoy & Amandi 2005].
La représentation d'un profil d'utilisateur est alors basée sur les concepts ontologiques qui modélisent les centres d'intérêts d'un utilisateur.

3.7.4 Filtering

Le filtrage consiste à filtrer les informations et à éliminer toutes les informations non pertinentes par rapport aux intérêts de l'utilisateur. La source du processus de filtrage peut être un ensemble de documents, par conséquent les documents seront filtrés selon leurs similarités aux profils, et les documents ayant le score le plus élevé seront présentés aux profils correspondants [Sheth 1994]. Les retours des utilisateurs sur les documents, qui peuvent être négatifs ou positifs, ont une influence sur la représentation du profil ainsi que sur les prochaines propositions de documents [Sheth 1994]. Les méthodes de filtrage sont «rule-based», «content-based», «collaborative-based» et hybride :

1. Rule based filtering :
   Le filtrage du flux d'informations entrant se fait dans ses approches sur la base des règles sous la forme de «if then»,

et sont spécifiées par le système d'information. L'utilisateur, à travers les étapes de son enregistrement, fournit directement un contenu d'informations démographiques qui seront utilisées par la suite pour la construction des règles [Choi & Han 2008]

2. Content based filtering :
   Ou appelé aussi «cognitive filtering». Dans ses approches, le filtrage s'effectue sur la base du contenu [Cufoglu 2014]. Cette forte dépendance au contenu présente l'avantage d'une analyse objective sans intervention de l'expert, mais elle présente un inconvénient en termes de volume. Dans ce cas, un volume important du contenu doit être fourni pour que le filtrage puisse donner un résultat pertinent. L'analyse du contenu peut être lié à des approches sémantiques tel que le Vector space model qui est une modélisation algébrique d'un document et qui considère sa dimension sémantique. Seuls les termes significatifs sont pris en compte. Chaque terme possède un poids qui représente son occurrence dans un document. Les requêtes, les concepts, les documents, sont tous considérés comme des vecteurs dans l'espace vectoriel [Wong & Raghavan 1984]. Le traitement d'une requête est alors basé sur la comparaison des vecteurs documents [Martinet *et al.* 2002]. Le latent semantic indexing, ou en français L'indexation sémantique, [Deerwester *et al.* 1990] permet d'analyser le contenu et par conséquent de le filtrer selon une matrice d'occurrence de terme «tronqués à la racine. Une matrice sera alors décomposée par des valeurs singulières (SVD). Si W est la matrice originale : W = T * S * D où T est la matrice de terme, S est une matrice diagonale de valeurs singulières et D est une matrice de documents. Cette représentation réduit la taille de la matrice principale et facilite le

calcul des similarités. Deux termes sont considérés similaires s'ils sont utilisés dans le même contexte, et deux contextes sont similaires s'ils comportent des mots similaires. Une requête est représentée par un vecteur de mots-clés, aussi appelé pseudo-code [Deerwester *et al.* 1990].

3. Collaborative based filtering :
   Le filtrage collaboratif établit un lien entre le comportement de l'utilisateur au passé avec son futur. Il suppose que si on connait les intérêts de l'utilisateur à un moment précis nous pourrons prédire ses intérêts futurs. Le collaborative filtering est considéré comme l'approche la plus pertinente pour la recommandation [Godoy & Amandi 2005]. L'idée est de filtrer le flux de données selon les centres d'intérêt identifiés pour des utilisateurs similaires. Par exemple, le système Ringo [Shardanand & Maes 1995] est un système de recommandation musicale qui propose à un utilisateur des albums et des chanteurs qui ont été bien notés par des utilisateurs similaires. L'étape de départ est d'identifier la cible d'utilisateurs. Ce mécanisme est utilisé par Amazon [Godoy & Amandi 2005] pour des recommandations qui apparaissent sous la forme de «les utilisateurs qui ont acheté le produit X ont aussi acheté le produit Y».
4. Hybride filtering :
   Les approches hybrides combinent des algorithmes différents de filtering. Suite à cette combinaison, les approches hybrides bénéficient des avantages de l'union des différentes techniques de filtrage. Un exemple d'utilisation de ses méthodes hybrides est l'approche PTV [Smyth & Cotter 2001] qui utilise une approche hybride pour recommander des programmes TV basés sur les profils d'utilisateurs représentés par leurs chaînes, mots clés, programmes. Ces profils sont

mis à jour selon les pertinences des retours. Cette approche prend en compte d'une part les intérêts d'un utilisateur et, d'autre part elle leur propose des recommandations selon les retours des profils similaires.

3.7.5 Statistical modeling

«Statistical modeling» est une technique de profilage qui consiste à construire des profils à partir de mots clés ou de logs [Eke *et al.* 2019]. Sur le web, cette technique consiste à représenter un profil d'utilisateur en se basant sur les termes les plus fréquents obtenus à partir des pages web visités par l'utilisateur. [Chen *et al.* 2010] ont par exemple présenté un modèle de recommandation d'URL pour recommander à un utilisateur de Twitter l'URL du flux de contenu qui pourrait l'intéresser. Les informations capturées depuis le profil Twitter englobent ses tweets, ses abonnés, les URLs de ses abonnés favoris, etc. Après un calcul de fréquence, elles seront capables de représenter son profil ce qui facilitera la personnalisation des propositions d'URL.

3.8 Discussion

Les techniques de profilage permettent d'établir une personnalisation de services ou de contenu en se basant sur les caractéristiques et le comportement d'un utilisateur. Cette personnalisation peut se faire de plusieurs façons : manuelle, automatique, implicite ou explicite. Il existe une diversité d'approches permettant d'appliquer un profilage statique ou dynamique.
Les dimensions à considérer dans notre adaptation d'une technique de profilage sont les suivantes :

— L'automatisation : afin de minimiser l'intervention des collaborateurs de l'entreprise, les approches de profilage automatique sont plus favorables ;
— Les techniques hybrides : le profilage hybride est un garant de qualité et de précision des profils. Le travail de profilage peut être fait d'une façon implicite en création et en mise à jour. Il peut être enrichi par des informations que l'expert ou le collaborateur souhaite compléter.
— Le filtrage hybride : les techniques de filtrage hybride combinent, d'une part, une comparaison d'un flux d'informations à celle d'un profil basé sur l'analyse du contenu (content based filtering), et d'autre part une comparaison de similarité de comportement d'un utilisateur avec d'autres afin de proposer les intérêts des utilisateurs jugés similaires.
— Le profilage de collaborateur : plusieurs informations des collaborateurs peuvent être collectées grâce au profilage. Dans le cadre de notre travail nous nous intéressons à certaines informations parmi d'autres. Ces informations, une fois extraites, permettent de répondre à deux questions importantes : quelles connaissances détient le collaborateur ? de quelles connaissances a-t-il besoin ? Pour pouvoir répondre à ces deux questions, un profil de collaborateur doit considérer deux axes : des informations statiques telles que : son affiliation et sa position dans l'organisation, et des informations dynamiques liées à son comportement, ses tâches et les projets auxquels il a participé (Figure 3.5).

3.9 Conclusion

Le profilage s'avère être une démarche qui a la capacité de répondre à certains points évoqués dans notre problématique de

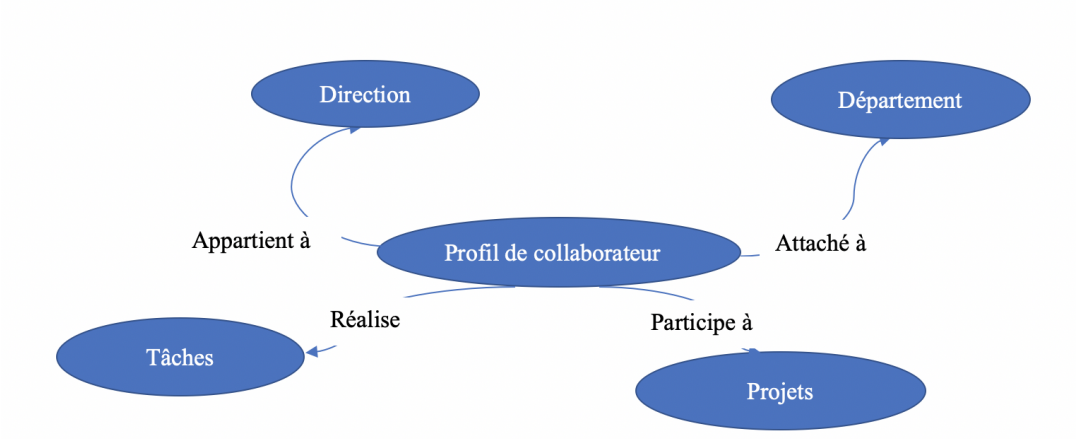


FIGURE 3.5 – Exemple de modélisation d'un profil

recherche détaillée dans le chapitre introductif. Nous citons ces points comme suit :

— Construire une approche basée sur les profils de collaborateurs : le collaborateur est le pilier de l'activité de l'entreprise. À l'aide du profilage nous pourrions représenter le profil de ces collaborateurs et comprendre leur implication dans la mémoire organisationnelle de l'entreprise. Évidemment cela permettra de tracer les périmètres de production et de besoin en connaissances de chaque collaborateur de l'entreprise ;

— Réduire le bruit : la réussite de la construction des profils précis et bien structurés permet d'établir un lien entre le collaborateur et les sources de connaissances qui l'intéressent, ce qui éliminera les informations non pertinentes et non intéressantes pour chaque profil ;

— Réduire le temps de recherche de la connaissance : le lien entre le collaborateur et les ressources qui l'intéressent élimine les ressources non pertinentes et permet un accès facile, rapide et direct aux supports de connaissances pertinents.

Notre approche Know-linking assure un profilage implicite et automatique qui, comparé à d'autres approches, réduit le temps et l'implication de l'utilisateur (voir tableau 3.2 ci-dessous).

TABLE 3.2 – Comparaison de Know-linking avec d'autres méthodes de profilage.

Approche	Implicite	Explicite	Automatique	Manuelle
Statistical keyword analysis [Schiaffino & Amandi 2000]		X	X	
Social filtering algorithms [Schiaffino & Amandi 2000], [Ismail *et al.* 2018]	X		X	
Machine learning techniques [Van Otterlo 2013]	X		X	
Rule based filtering [Kuflik *et al.* 2003]		X	X	
Collaborative filtering [Stefanidis *et al.* 2018]	X		X	
Content based filtering [Geetha *et al.* 2018]	X		X	
Questionnaires [Klement 2015]		X		X
Know-linking [Abderrahim *et al.* 2020]	X		X	

Nous proposons donc d'intégrer une technique de filtrage hybride et automatique afin de construire des profils de collaborateurs. Cela permettra d'enrichir les principes de notre approche Know-linking pour assurer une traçabilité et un partage de connaissances personnalisé et automatisé (figure 3.6).

L'étude de profilage nous a apporté une réponse sur la manière de procurer un cadre de personnalisation à notre approche et en même temps elle a offert de nouvelles pistes de réflexion à prendre

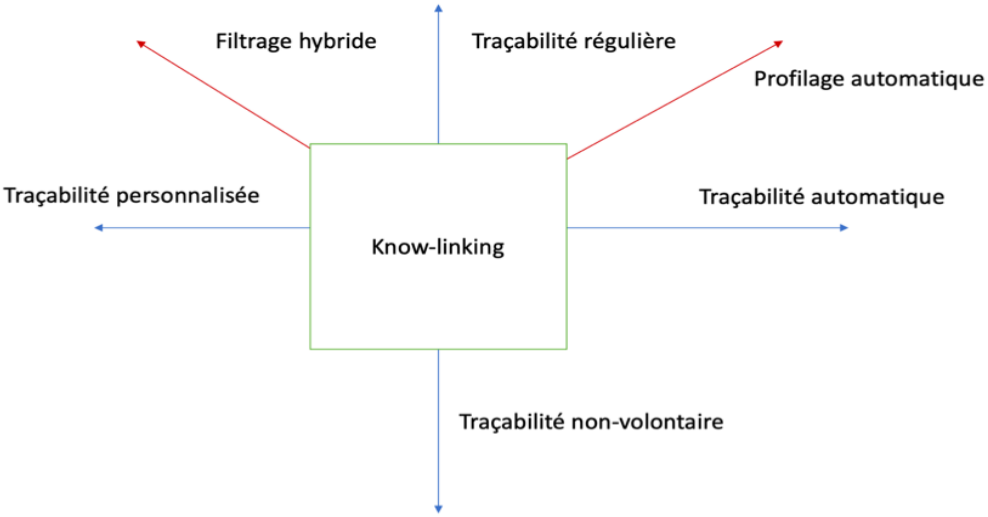


FIGURE 3.6 – Nouvelles dimensions de Know-linking

en considération pour la suite de notre travail comme par exemple : existe-il des liens entre les profils de collaborateurs ? Et comment pourrions-nous analyser le contenu textuel pour organiser les documents selon les profils ? Nous explorons les approches en lien avec ces questions dans le chapitre suivant.

Chapitre 4

Analyse du contenu et organisation des documents

«THE LIMITS OF MY LANGUAGES ARE THE LIMITS OF THE WORLD" Ludwig Wittgenstein

Sommaire

4.1 Introduction

Les supports documentaires sont la source la plus importante de la connaissance d'aujourd'hui. D'après les études de l'entreprise NINTEX [NINTEX], ces documents sont générés par plusieurs sources : ils sont créés directement par les employés ou bien générés par les systèmes d'information tels que les rapports et les bilans. Ces documents peuvent être électroniques (PDF, documents Word ou Excel..) ou bien constituer des supports physiques (papier). Procéder à l'analyse de ces documents permet d'enrichir le patrimoine de la connaissance de l'entreprise par de nouvelles connaissances explicites. En examinant la littérature scientifique sur le sujet en réponse à notre réflexion «comment pourrions-nous analyser le contenu textuel ?», nous trouvons plusieurs méthodes permettant de l'analyser. Nous consacrons la première partie du présent chapitre à la présentation des différents travaux dans ce contexte. Bien que l'analyse du contenu textuel soit un champ prometteur pour extraire et capitaliser des connaissances rares et variées, il faut s'investir à rendre ces blocs de textes jugés importants accessibles aux différents usagers. Cela met l'accent sur un autre champ de recherche complémentaire à celui de l'analyse du contenu, celui de l'organisation des documents, un domaine qui englobe l'indexation, la catégorisation et la classification des documents. Nous consacrons la seconde partie de ce chapitre à détailler les efforts de recherche présentés dans la littérature dans le contexte de l'organisation des ressources documentaires de l'entreprise. Nous exposons comment ces travaux peuvent être utilisés et adaptés dans notre approche «Know-linking».

4.2 Préparation du contenu textuel pour l'analyse

Généralement, dans un texte, environ 20% à 30% de la totalité des mots sont des mots de ponctuation [Kannan *et al.* 2014]. Ainsi, le contenu textuel à son état brut n'est pas prêt pour une analyse vu qu'il contient beaucoup de «bruit» qui impactent étroitement la qualité des résultats de la méthode de l'analyse adoptée. La phase d'élimination du bruit et le nettoyage du contenu est nécessaire avant de procéder au traitement. Cette phase est appelée «prétraitement». Le prétraitement du texte est un travail préliminaire. Nous détaillons les différents exercices de préparation du texte dans les points suivants :

— Tokenization : technique fondamentale pour la majorité des tâches de Traitement Automatique de Langage Naturel TALN [Solangi *et al.* 2018]. La tokenization [Jettakul *et al.* 2018] consiste à segmenter (en anglais «split»), un document ou bien des blocs de texte sous forme de «tokens» qui peuvent être des expressions ou des mots. En langue française ou anglaise par exemple les mots sont séparés par des espaces. On y ajoute certaines connaissances linguistiques pour éliminer les mots pas très important à analyser comme en français les articles 'le ', 'la ','les' ou ' un', 'une'.. ou en anglais 'the' ou bien 'a'. Cette phase de nettoyage et de préparation élimine la ponctuation et les autres séparateurs du texte. Cette tâche est fondamentale et il existe plusieurs outils permettant la tokenization. On peut citer par exemple Stanford Tokenizer [Solangi *et al.* 2018], OpenNLP Tokenizer [Dumal *et al.* 2017].

— POS tagging [Straka & Straková 2017] : Part of speech tagging, ou en français «étiquetage morphosyntaxique». Cela

consiste à identifier la classe morphosyntaxique qui est associée aux mots dans leur contexte d'énonciation. Le «pos-tag» désigne la classe morphosyntaxique d'un mot qui peut être un adjectif, un verbe, un nom, etc. L'objectif de cette étape est donc de déterminer les informations lexicales concernant les mots d'un texte. [Schreiber *et al.* 2018]

— Parsing : cette étape du prétraitement du texte consiste à hiérarchiser les mots d'une phrase. Le Parsing [Jain *et al.* 2020] représente la structure grammaticale d'une phrase donnée sous forme d'un arbre. L'arbre de la phrase représente les relations qui existent entre les mots constituant cette phrase.

4.3 Analyse du contenu textuel

Le terme «contenu» au sens large englobe les informations stockées dans des supports de différents formats que ce soient des vidéos, des enregistrements vocaux ou des documents (électroniques ou papier), autrement dit «tout ce qui est dit ou écrit» [Henry & Moscovici 1968]. L'importance de ces supports a suscité des champs de recherche qui visent à extraire les connaissances et les traces à partir de ce contenu varié et à les rendre facilement accessibles. Paul Henry et Serge Moscovici [Henry & Moscovici 1968] pensent que l'analyse du contenu «est un ensemble disparate de techniques utilisées pour traiter des matériaux linguistiques». Dans nos travaux, nous nous intéressons plus à l'analyse du contenu textuel puisque celui-ci représente la source la plus importante de la connaissance de l'entreprise.

4.3.1 Le traitement des langages naturels

Il est important de distinguer le langage formel du langage naturel avant d'aborder le sujet du traitement de la langue. La langue

naturelle est un langage développé naturellement sans planification ou modélisation [Yvon 2010]. Autrement dit, elle désigne la langue «parlée ou «écrite» par les êtres humains, par opposition aux langages formels : artificiels, informatiques, mathématiques ou logiques [Arbi 2014] tels que le français ou l'anglais. Ainsi ce langage est dit humain, d'où l'adjectif naturel, et non pas langage formel [Yvon 2010].

Le langage formel est le langage de la machine et il est souvent utilisé afin de programmer la machine à exécuter certaines tâches, autrement dit c'est un canal de communication homme-machine. Cependant, le langage naturel est le langage que les êtres humains utilisent pour communiquer entre eux. Le traitement automatique des langues naturelles est donc dédié à la conception de méthodes et d'outils informatiques pour analyser la langue humaine [Kessler 2009].

D'une façon plus précise, le traitement automatique du langage naturel «TALN» est un domaine de l'informatique qui s'intéresse à l'interprétation et à la production par des «machines de phrases» ou de textes dans des langues telles que le français ou l'anglais [Gilloux 1989]. Cette capacité d'interprétation et de l'analyse du texte n'est pas une informatisation d'un processus humain mais plutôt un travail automatisé et autonome de la machine. Dans ce sens, certains travaux de recherche définissent le traitement automatique de langages naturels comme un domaine de recherche et aussi d'application qui explore «comment les ordinateurs peuvent être utilisés pour comprendre et manipuler un discours ou un texte écrit en langage naturel» [Chowdhury 2003]. En conclusion, nous pouvons noter que le traitement automatique des langages naturels présente un espace d'intersection de deux domaines : l'informatique et la linguistique.

Les tâches que le «TALN» cherche à résoudre sont relatives aux

différentes branches de la linguistique telles que la syntaxe (lemmatisation, étiquetage morphosyntaxique, analyse syntaxique, etc.), la sémantique (reconnaissance d'entités nommées, traduction automatique, génération automatique de textes, etc.), le discours (résumé automatique, résolution de coréférence, analyse du discours, etc.) et le traitement de la parole (reconnaissance automatique de la parole, synthèse vocale, etc.) [Dalloux 2020]. L'objectif de ce nouveau champ (Le TALN) est donc de confier aux ordinateurs des tâches utiles impliquant le langage humain comme permettre la communication homme-machine et améliorer la communication homme-homme, ou bien tout simplement traiter un texte ou un discours [Jurafsky & Martin 2013].

L'aspect de la modélisation est aussi présent dans les définitions du TALN dans certains travaux de recherche. Par exemple, [Liddy 2001] définit le TALN comme des techniques informatiques permettant de représenter et de comprendre le langage humain. François Yvon quant à lui considère le TALN comme un «ensemble des recherches et développements visant à modéliser et à reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication.» [Yvon 2010].

Le traitement automatique des langages naturels est en résumé un domaine qui vise à réduire le gap entre le langage humain ou naturel et la capacité de la machine à comprendre et analyser ce langage. Cette intelligence fournie à la machine peut être développée de différentes manières. Dans la littérature scientifique on retrouve deux catégories de TALN : on distingue notamment des approches syntaxiques liées à la théorie des langages formels et d'autres approches numériques qui s'appuient sur la probabilité et les statistiques [Kessler 2009].

4.3.1.1 Les méthodes du TALN

[Cori 2008] a distingué dans ses études des outils du TALN [Cori 2008] les méthodes du traitement automatique de langue «robustes» et d'autres méthodes «théoriques» en se basant sur trois critères. Le premier est la qualité des données linguistiques : l'auteur distingue les données venant de «vraies» productions langagières et les données fabriquées par les linguistes. Le second critère est le non-blocage du système sous prétexte que les données seraient «incorrectes» ou non grammaticales. Le troisième critère est celui de l'évaluation (quantitatives) de la performance [Cori 2008]. Les méthodes du TALN «robustes», d'après l'étude de [Cori 2008], sont classées en deux familles : les méthodes à base des modèles probabilistes/statistiques, ou aussi appelées stochastiques, et les méthodes basées sur des règles. Les méthodes stochastiques sont des méthodes probabilistes fondées sur des calculs statistiques effectués à partir de corpus [Cori 2008]. Une méthode générique qui s'applique à presque toutes les tâches du TALN stochastique est présentée par Bernard Merialdo [Merialdo 1995]. Cette méthode générique se compose à son tour de trois phases principales.
Premièrement, la modélisation du problème qui exige l'identification d'une hypothèse. Par exemple, on suppose que «l'apparition d'un mot ne dépend que des deux mots précédents».
Deuxièmement, la construction des estimations, une étape durant laquelle on construit des estimations des valeurs des «probabilités élémentaires» (définies dans la phase de modélisation). Enfin, l'application de la probabilité élémentaire sur des données nouvelles. Dans le contexte des méthodes du TALN stochastique il existe d'autres méthodes telles que la correction orthographique [Kemighan *et al.* 1990] qui repose sur la correction de l'orthographe des mots indépendamment du contexte. Dans le même cadre s'inscrit

la méthode des n-grammes qui se base sur un calcul de probabilité afin de déterminer la probabilité d'apparition d'un mot à la suite des autres, ce qui permet de prédire l'apparition d'un mot dans une suite. Cependant, les méthodes basées sur des règles ne mettent pas en jeu de comptages. Elles sont utilisées par exemple pour étiqueter des textes, c'est-à-dire pour affecter une catégorie, ou partie du discours, à chacun des mots qui le composent» [Cori 2008]. Les méthodes basées sur des règles utilisent des règles linguistiques, ou aussi dites contextuelles, pour enlever l'ambiguïté. En mettant le mot dans son contexte, on peut reconnaître s'il s'agit bien d'un nom, d'un verbe, d'un article, etc. Le type d'ambiguïté que ces méthodes essayent de résoudre concerne par exemple les mots dit homonymes nom/verbe dont l'identification est difficile sans connaître le contexte de leur emploi, par exemple avec les mots «ferme», «accusé» ou encore «avion». Par conséquent, on ne peut pas les analyser facilement. Un exemple de règle employée pour une telle problématique est la suivante : si par exemple le mot «accusé» est précédé d'un article comme «un», «accusé» est donc un nom.

Des années plus tard, [Cambria & White 2014] ont travaillé sur une étude des méthodes et outils du TALN en mettant en valeur l'évolution des méthodes du TALN vers la sémantique. Cette étude positionne la nécessité des approches sémantiques par rapport aux trois familles de méthodes principales qui représentent, d'après les auteurs, l'évolution du TALN. Ces méthodes sont respectivement les méthodes syntaxiques (Syntactics) [Harris 1954] telle que la méthode «bag-of-words», sémantique (Semantics) [Cambria] telle que la méthode «bag-of-concepts», (cette méthode résout essentiellement le problème de l'ambiguïté des mots synonymes), et pragmatiques (Pragmatics) telle que la méthode «bag-of-narratives model» qui offre une représentation de chaque pièce d'un texte sous

forme d'une mini-histoire (en anglais «mini story»), ou bien des épisodes interconnectés, ce qui permet d'avoir un niveau plus détaillé au niveau de la compréhension du texte. Par exemple, la résolution des co-références : dans une phrase comme «le professeur a demandé aux élèves de faire l'exercice. Il a recommandé un outil pertinent», la résolution de co-référence consiste à détecter que le pronom «il» fait référence au professeur.

La syntaxe spécifie alors la façon dont les groupes de symboles doivent être organisés, de sorte que le groupe de symboles soit considéré comme correctement formé. La sémantique spécifie ce que les expressions bien formées sont censées signifier. La pragmatique spécifie comment les informations contextuelles peuvent être exploitées pour fournir de meilleures corrélations entre différentes sémantiques, ce qui est essentiel pour des tâches telles que la désambiguïsation du sens des mots [Cambria].

Ces techniques viennent en complément des méthodes syntaxiques qui se contentent d'analyser ce qui est «vu» comme la fréquence d'apparition des termes, ce qui représente une limite face aux différents sens et rapports qui existaient entre les termes dans un texte.

Certains travaux de recherche se sont concentrés sur la réduction du gap cognitif entre la capacité humaine à analyser un texte et celle d'une machine, autrement dit à améliorer la capacité d'une machine à comprendre le sens des termes. Parmi les approches du TALN sémantique les plus utilisées on retrouve l'analyse sémantique latente, une méthode qui considère les relations entre les termes d'un document. La technique de l'analyse sémantique latente (LSA) cherche à construire un espace sémantique de très grande dimension en se basant sur une analyse statistique de l'ensemble des cooccurrences dans un corpus de textes [Deerwester *et al.* 1990], [Bestgen 2006]. L'idée de LSA est de construire une

matrice d'occurrence de termes dont les lignes représentent les termes et les colonnes les documents qui les contiennent, un document pouvant être un texte, un paragraphe ou même une phrase [Bestgen 2006], tout en se basant sur la pondération (en anglais «term frequency») que le nombres d'apparition des termes est normalisé [Bestgen 2006]. Dans l'approche [Deerwester *et al.* 1990], le sens de chaque mot y est représenté par un vecteur. Afin de pouvoir mesurer la «similarité sémantique» entre deux termes différents, on calcule le cosinus entre les vecteurs qui les représentent [Bestgen 2006]. Cette méthode du TALN a été étendue en ajoutant un modèle statistique particulier pour donner l'analyse sémantique latente probabiliste (en anglais «Probabilistic latent semantic analysis») [Hofmann 1999]. Etant inspiré de l'analyse sémantique latente, comme son nom l'indique, tout document d'une collection D est représenté par une distribution de probabilité sur les K valeurs de la variable thématique latente A = 1,..., K, où chaque valeur de correspond à une distribution de probabilité sur l'ensemble des mots de la collection. Un document est d'abord choisi suivant la probabilité P(d), ensuite une thématique est générée avec une probabilité P(|d), et finalement un mot w est émis suivant la probabilité P(w|) [Hofmann 1999].

La PLSA est souvent appliquée dans l'indexation et le filtrage d'information ainsi que dans l'apprentissage à partir du texte et dans d'autres domaines en lien avec l'analyse du texte.

L'évolution des méthodes du TALN, depuis l'étude de Erik Cambria, a été plus tard illustrée dans l'étude de Shiliang Sun Et al qui est orientée dans un domaine d'application plus spécifique : l'Opinion mining [Sun *et al.* 2017]. Cette capacité des méthodes du traitement des langages naturels à détecter un sentiment positif ou négatif à partir d'un bloc du texte. L'une des méthodes

appliquée est le calcul de polarité pour accorder une valeur positive à un terme considéré comme mélioratif et une valeur négative pour les termes péjoratifs. À la fin, la somme est calculée. Cette approche est améliorée par des algorithmes de classification qui sont entrainés à déterminer la polarité d'un texte, comme les algorithmes naïves bayes et support vector machine. On constate que les algorithmes d'apprentissage automatique «machine learning» peuvent améliorer la performance des méthodes du TALN.

4.3.1.2 Les outils du TALN

Il existe plusieurs outils sur le marché qui permettent de traiter automatiquement les langages naturels afin d'analyser des textes dans différentes langues. Ces outils peuvent être des logiciels, des bibliothèques ou des plateformes. Parmi les outils les plus utilisés du TALN on retrouve GATE (General Architecture for Text Engineering) [Tablan *et al.* 2004]. GATE est une infrastructure développée en 1995 à l'Université de Sheffield qui permet de développer et de déployer des composants pour le TALN. L'infrastructure GATE présente une architecture en plus d'un framework en Java et un environnement de développement intégré. La limite de Gate pour un linguiste est qu'il nécessite des connaissances en programmation (notamment en JAPE).

L'outil NOOJ [Silberztein 2016], est un environnement de développement linguistique et d'analyse de corpus autonome permettant de construire, de tester et de maintenir des descriptions formalisées à large couverture des langues naturelles, ainsi que de développer des applications du traitement de la langue. Le successeur de INTEX [Silberztein 1993], adopte un seul formalisme (modèle d'analyse) basé sur des automates. NOOJ est constitué de trois modules qui sont principalement : «corpus handling», «lexicon» et «grammar». NOOJ est limité en utilisation en dehors de

l'environnement Windows car il est développé en NET. Il existe des architectures logicielles pour le développement du TALN telle que UIMA (Unstructured Information Management Architecture). Cette architecture a pour objectif de décrire les étapes à suivre pour traiter automatiquement et extraire les informations à partir d'un texte, d'une image ou d'une vidéo. Cependant, cette architecture est considérée comme abstraite et les modules d'analyse du texte doivent être implémentés par le concepteur.
Le NLTK (Natural language toolkit) [Bird 2006] est un ensemble de module python open source permettant le prétraitement du texte (tokenization, le POS tagging) et le raisonnement sémantique. NLTK offre plusieurs types de données comme : tokens, tags, chunks, trees, et il fournit des interfaces pour de nombreux corpus et lexiques qui sont surtout utiles pour l'exploration d'opinions «opening mining» et l'analyse des sentiments. La bibliothèque Apache openNLP est une bibliothèque Java permettant de faire un prétraitement d'un texte ; elle supporte les tâches de TALN comme la reconnaissance d'entités nommées et la résolution des coréférences.

4.3.2 Fouille de texte (Textmining)

Les études ont démontré que 85% des informations résident dans les documents [Kumar & Bhatia 2013]. Cette importance du contenu textuel a permis de faire apparaître des approches pour fouiller les textes et en extraire les informations utiles. Le terme Textmining, ou fouille de texte, est apparu et a été cité pour la première fois dans Feldman et al [Feldman & Dagan 1995]. Le terme Text mining est employé pour désigner la découverte des connaissances à partir du texte (KDT) (knowledge discovery from text). Dans la littérature scientifique il existe trois courants du

textmining [Hotho *et al.* 2005]. Le premier considère le textmining comme un processus de KDD, le second le voit comme une technique d'extraction d'information. Enfin le dernier l'envisage comme un datamining appliqué aux textes.

— Le Textmining en tant que KDD Process : Le Textmining est vu comme tout processus composé d'un ensemble d'étapes incluant les procédures statistiques et les algorithmes de data mining. Gomez, dans son approche hidalgo2002tutorial, considére le textmining comme un processus orienté vers les approches textuelles.

— Le Textmining en tant que datamining : l'application des méthodes et des algorithmes des disciplines machine learning et statistiques sur des textes dans le but de trouver des patterns utiles. Pour cela il est nécessaire de bien préparer le texte avant de le traiter. Plusieurs approches consistent à appliquer le TALN ou de simples méthodes d'extraction pour extraire des données à partir du texte puis appliquer des algorithmes de data mining sur les données extraites.

— Le Textmining en tant qu'extraction d'information : dans cette approche, le textmining est supposé correspondre à l'extraction de l'information, des faits à partir du texte et l'interprétation sémantique et syntaxique [Hotho *et al.* 2005].

4.3.2.1 Approches et Techniques de textmining

Les approches de textmining ont connu une évolution importante ces dernières années. Ces approches présentent la base de plusieurs outils. Il existe des techniques considérées de base pour le textmining tel que le bag-of-words [Kim *et al.* 2017], qui consiste à modéliser le texte par une suite de mot (après prétraitement). Cette méthode peut être utilisée pour comparer des documents.

Deux documents sont ainsi comparés selon les mots qui les représentent, et le calcul de similarité se fait en fonction du nombre des mots en commun. Ce calcul de similarité est souvent appelé «coordinate matching» [Nahm & Mooney 2002].
Un terme dans un corpus peut avoir deux poids, le premier poids étant calculé selon son nombre d'occurrence dans un document, aussi appelé «term-frequency», et l'autre poids correspond au calcul de son nombre d'apparition dans tous les documents du corpus, en se basant sur le principe que les mots très fréquents «tf» portent moins d'informations que les mots rares, soit le «inverse document frequency», ou en abréviation «idf». Les deux poids calculés seront combinés par multiplication «tf * idf» afin de donner le principe d'une méthode largement utilisée en textmining appelée «td*idf» [Harish & Revanasiddappa 2017].
Les modèles probabilistes à états finis ou Modèles de Markov cachés (HMM) [Hotho *et al.* 2005] analysent grammaticalement, ou en anglais «parse», une séquence d'entrée en suivant son flux à travers le modèle, ce qui permet d'avoir une représentation de l'état actuel du modèle par une distribution de probabilité de tous les états. Généralement, l'état initial est inconnu ou «caché» et doit être à son tour représenté par une distribution de probabilité. Chaque nouveau «token» affecte cette distribution d'une manière dépendant de la structure et des paramètres du modèle. Finalement, la majorité des probabilités peut être concentrée sur un état particulier, ce qui permet de lever l'ambiguïté de l'état initial et même de toute la trajectoire des transitions d'états correspondant à la séquence d'entrée. Les part of speech taggers sont basés sur cette technique de HMM [Brill 1992].

4.4 Organisation des documents

Les mécanismes d'organisation des documents de l'entreprise reposent sur plusieurs principes qui permettent de rendre les documents accessibles et d'optimiser le temps de recherche d'un collaborateur intéressé par un contenu. Plusieurs mécanismes permettent de lier le collaborateur avec le contenu de l'entreprise, par exemple les systèmes de recommandation, l'indexation et la classification des documents.

4.4.1 L'indexation des documents

L'indexation est un processus permettant de construire un ensemble d'éléments «clés» afin de caractériser le contenu d'un document /et de retrouver ce document en réponse à une requête [Bouhriz *et al.* 2014]. Ce processus d'indexation vise à constituer une représentation du contenu des documents et des requêtes afin de procéder à un appariement pertinent entre eux [Bouhriz *et al.* 2014]. L'indexation en tant que discipline consiste à trouver une façon de représenter un contenu pour faciliter, par la suite, son identification. Ce processus permet de créer une «courte description ou caractérisation du contenu d'un document textuel sous forme de représentation suivant un modèle» [Moens 2000]. Ces modèles et techniques peuvent être mathématiques, donc l'indexation peut aussi être l'ensemble des techniques mathématiques permettant d'optimiser une recherche d'informations [Moens 2000].
La description du contenu textuel peut être indexée selon un vocabulaire qui peut être libre si l'indexeur décrit le contenu avec un langage naturel issu du document lui-même [Harter 1986], ou contrôlé par un ensemble de descripteurs figés et préétablis choisis en amont du processus d'indexation.
Revenons à l'indexeur qui peut être une personne (expert) ou un

mécanisme (une machine ou algorithme) qui se charge du processus d'indexation. Ce processus révèle trois types différents d'indexation, à savoir l'indexation manuelle si l'indexeur est un être humain (expert), ou bien l'indexation automatique si c'est l'ordinateur qui le prend en charge. Les deux types d'indexation peuvent être combinés pour donner l'indexation semi-automatique, autrement dit effectuée par une machine sous la surveillance d'un être humain.

4.4.1.1 Les techniques d'indexation

Les techniques d'indexation ne sont pas loin de l'évolution rapide des approches informatiques et de l'intelligence artificielle. Ces techniques peuvent être classées en trois grandes familles [Gani *et al.* 2016] : les méthodes traditionnelles ou classiques de l'indexation, les méthodes basées sur l'intelligence artificielle et la combinaison des deux qui forme l'intelligence artificielle collaborative. Les méthodes classiques sont une famille qui regroupe les algorithmes d'indexation non-intelligence artificielle, par exemple l'approche bitmap.

Le bitmap [Chan & Ioannidis 1998] est une technique d'indexation basée sur des données dites données d'indice en bloc enregistrées sous forme de séquences de bits. Les séquences de bits sont utilisées dans les opérations logiques pour répondre aux requêtes. Ainsi, la représentation sous forme de graphes peut servir à la description du contenu textuel. Dans cette catégorie on retrouve l'approche des graphes [Sakr & Al-Naymat 2010]. La base de cette approche est la modélisation du graphe, un supergraph ou subgraph, qui consiste à récupérer tous les graphes de la base de données de telle sorte qu'un graphe de requête donné en soit un sous-graphe, puis indexer tous les arcs du graphe et leurs occurrences. Parmi les approches de la deuxième famille on retrouve la représentation sémantique ou ontologique [Gani *et al.* 2016]. Des approches

offrant une modélisation du contenu textuel avec un aspect de raisonnement. L'approche sémantique propose d'annoter le texte. Chaque annotation de chaque document est stockée dans une base de connaissances et une pondération est attribuée pour refléter la pertinence de l'entité ontologique par rapport à la signification du document. L'idée principale est que plus les concepts sémantiquement proches apparaissent dans un document, plus les valeurs de vecteur obtenues par ces dimensions de concept sont élevées. L'indexation sémantique peut servir de base à des processus de recherche améliorés pour des masses volumineuses de données. Dans la troisièmes famille de méthodes d'indexation on combine deux méthodes de deux familles différentes, par exemple le «collaborative knowledge representation and raisoning» CKRR, basé sur la représentation collaborative (à l'aide des tags) des connaissances. Cette méthode est développée sur la base de l'apprentissage social en collaboration avec la représentation des connaissances afin de proposer une solution d'indexation collaborative permettant une extraction efficace des documents liés sémantiquement. Les méthodes d'indexation, qu'elles soient basées sur les algorithmes de l'intelligence artificielle ou la représentation sémantique, sont la base de plusieurs travaux de recherche en indexation.
Nous comparons quelques travaux de la littérature dans le tableau suivant 4.1.

TABLE 4.1 – Tableau comparatif des approches d'indexation.

Approche	Auteurs	Méthodologie	Type d'indexation	Observation
Top-K queries on temporal data [Li *et al.* 2010]	Li et al	-Mapping données /mots clefs -La méthode d'indexation basée sur B-Tree trace une collection d'échantillons p pour une série de valeurs de données temporelles décroissantes géométriquement -Il faut un temps quasi-linéaire pour répondre à toute requête top-k (t) avec le coût d'E/S optimal souhaité	B-Tree	-Temps de réponse considérable -La relation entre les termes n'est pas prise en considération
Indexing in network trajectory flows [Popa *et al.* 2011]	Sandu Popa et al	Méthode hybride (Graph + B-Tree)	Composite tree (B-Tree)	-Opération couteuse -Temps de réponse considérable -Relation entre les termes

Approche	Auteurs	Méthodologie	Type d'indexation	Observation
Fast graph query processing with a low-cost index [Cheng *et al.* 2011]	Cheng et al	- Basée sur la théorie des graphes - L'indice est construit par l'extraction des points communs entre les graphes. Traitement de requête a deux techniques clés -l'inclusion directe et le filtrage. L'inclusion directe permet d'inclure directement des réponses partielles à une requête sans vérification du candidat -Technique de filtrage réduit davantage l'ensemble de candidats en opérant sur une base de données projetée beaucoup plus petite.	Graph query tree	- Essentiellement pour résoudre la recherche dans les subgraph -Complexité des graphes
Creating a semantically enhanced cloud service environment through ontology evolution [Rodríguez-García *et al.* 2014]	Rodirguez garcia et al	-Chaque annotation est stockée dans une base de données avec un poids. -Un vecteur de relation entre concepts est établi	Semantic annotation	- Plus de pertinence sur les résultats - Problèmes avec les données volumineuses -Temps de réponse à la requête
Efficient storage of healthcare data in xml-based smart cards (armagan) [Gündem & Armağan 2006]	Gundem et al	- Conversion en séquence de bit - Applications logiques sont appliquées. Pour répondre aux requêtes	bitmap	- Simple - Non-coûteuse - Absence de sémantique

Approche	Auteurs	Méthodologie	Type d'indexation	Observation
Authentification of lossy data in body-sensor networks [Cheng *et al.* 2011]	Cheng et al	- On applique une fonction de hachage pour les données et les requêtes - Comparaison de similarité	hashing	- Facilite les opérations de recherche - Pas de traitement sémantique
iSAX : Indexing and Mining Terabyte Sized Time Series [Shieh & Keogh 2008]	- Jin Shieh et al	- Les mots sont représentés suivant. La technique Sax et les requêtes aussi - Les représentations sont comparées et une fois deux représentation sax sont égales le document contenant est retourné	multi- resolution symbolic representation	- Efficace pour chercher dans un volume important de données. - Les aspects sémantiques ne sont pas pris en compte
Building and using a medical ontology for knowledge management abd cooperative work in a health care network [Dieng-Kuntz *et al.* 2006]	Dieng-Kuntz	- Ils associent les caractéristiques du système de recommandation aux concepts sémantiques contenus dans les vidéos. - Les préférences utilisateur du système de recommandation sont appliquées lors de la recherche de vidéos similaires pour le concept d'utilisateur.	collaborative-based semantic	- Indexation à base sémantique - Base sur une ontologie

Approche	Auteurs	Méthodologie	Type d'indexation	Observation
Collaboration-based medical knowledge recommendation [Huang *et al.* 2012]	Huang et al	- Traduire une base de données médicales en RDF - Pour les données textuelles un traitement de langage naturel est appliqué - L'outil Virtual Staff a été créé pour assurer des diagnostiques par les membres. - Un système de filtrage est mis en place, il extrait les connaissances médicales les plus pertinentes en utilisant des techniques de recherche de mots clés	Collaborative learning	- Automatisé - Supervisé - Bénéficie des avantages du traitement des langages naturels - Le concept de profil de confiance n'est pas évident

4.4.2 Classification des documents et techniques d'apprentissage automatique

La classification des documents est un domaine de recherche consistant à attribuer des sujets à une collection de documents. Des algorithmes semi-supervisés sont appliqués pour identifier automatiquement une étiquette sur de nouveaux documents en fonction de l'analyse du texte. Une première étape de prétraitement est nécessaire pour préparer le texte à analyser. Cette étape consiste à supprimer les mots inutiles comme les auxiliaires et les articles. Dans la littérature scientifique, on retrouve des représentations différentes de documents. Certaines d'entre eux sont basées sur des représentations d'espaces vectoriels [Ikonomakis *et al.* 2005], qui représentent un document sous la forme d'un tableau de mots,

et l'ensemble de tous les mots d'un ensemble d'apprentissage est considéré comme du vocabulaire. On attribue la valeur 1 si le document contient le mot, sinon un 0 est attribué. Cette représentation binaire ne prend en compte aucune dimension sémantique des mots, ce qui impacte fortement les résultats donnés. D'autres méthodes sont basées sur la sélection d'entités qui vise à réduire la dimensionnalité du jeu de données en supprimant les entités considérées comme non pertinentes pour la classification [Forman 2003]. Par exemple, Guan et Zhou ont proposé une approche basée sur l'élagage du corpus de formation afin d'accélérer le processus [Guan & Zhou 2002]. L'idée principale de leur approche est de réduire de manière significative la taille du corpus de formation tant que les performances de classification peuvent être maintenues à un niveau de classification approximativement identique ou proche sans élagage du corpus de formation. Une variété d'algorithmes d'apprentissage automatique sont appliqués dans les classifications de documents, par exemple Naïve Bayes en raison de sa simplicité ainsi que de son efficacité. Le point faible des Naïves bayes est la modélisation du texte car elle impacte ses performances [Kim *et al.* 2002]. Schneider, dans [Schneider 2005], a représenté une solution pour corriger la modélisation du texte dans les bayes naïves. La méthode kNN de classification de texte a également été améliorée en utilisant des paramètres bien estimés. Ces approches ne prennent pas en compte la dimension sémantique des mots parce qu'un texte n'est ni une représentation binaire ni une valeur numérique qui apparaît ou non dans une séquence de codes numériques. Les liens entre les mots peuvent changer tout le contexte dans un texte, ce qui a un impact important sur les résultats de la classification. Cela a motivé les chercheurs à travailler sur la classification sémantique de textes. La classification sémantique est classée principalement en cinq catégories à savoir : méthodes

basées sur les corpus (analyse sémantique latente) de domaine, méthodes d'apprentissage en profondeur (deep learning), approches des séquences de mots / caractères améliorées et approche linguistique enrichie (règles lexicales et syntaxiques pour l'extraction des phrases nominales).

4.5 Discussion

Les approches du traitement de langage naturel et celles d'indexation ou de classification de documents peuvent ajouter certaines dimensions à notre approche :

— Comprendre automatiquement le contenu d'un document : la compréhension automatique et l'analyse du document permettent d'identifier le sujet d'un document, un travail préliminaire pour reconnaître la personne qui a besoin de ce document.

— Modéliser le contenu textuel : les techniques d'analyse du contenu permettent de modéliser le contenu textuel selon une représentation comme les graphes sémantiques.

— Lier les collaborateurs et le contenu : l'absence de ce lien laisse la place à plus de bruit et à plus de temps de recherche. Pour cette raison, l'analyse du contenu doit être performante et couvrir des aspects d'analyse sémantique. L'organisation des documents doit être à la base de ce lien de similarité entre le besoin du collaborateur et le contenu du document.

— Faire évoluer et enrichir automatiquement le capital de connaissances de l'entreprise : de nouvelles connaissances peuvent être extraites à l'aide des outils du TALN et du Textmining. On peut capter l'évolution dans l'entreprise en termes de processus, thématiques, outils, ou même organisationnelle par l'apparition de nouveaux concepts dans les documents

de travail. L'analyse de ces documents et l'extraction automatique de ces nouveaux concepts va permettre d'avoir une approche de partage de connaissances enrichie continuellement.

4.6 Conclusion

Les analyses présentées dans ce chapitre offrent une réponse à deux réflexions : la première est une question souvent très posée dans un organisme, «Qui a besoin de ce document ?». L'approche permettant d'avoir une réponse à cette question est composée de deux parties :

1. L'identification sémantique du contenu d'un document :
   Dans une entreprise où le nombre de documents produits au quotidien est important, les techniques de l'analyse du contenu présentent un élément clé pour interpréter et comprendre automatiquement le contenu d'un document. Une fois ce contenu interprété on pourra procéder à l'identification des personnes dont ce document fait partie de leurs intérêts et dans ce cas on aborde le sujet de l'organisation des documents qui explique la logique ou la stratégie de rendre les documents accessibles aux collaborateurs.
2. La traçabilité des connaissances :
   Ce travail qu'on cherche à mettre à jour régulièrement et automatiquement est possible à l'aide des algorithmes du TALN et du Textmining. En effet, une application continue de ces techniques permet d'enrichir les liens entre les documents et les besoins des acteurs dans l'entreprise et de mettre ainsi en avant les évolutions des thématiques et des organisations dans cette dernière.

Nous modélisons toutes les dimensions à considérer et les options de notre approche Know-linking en prenant en compte l'avantage que peut offrir l'analyse du contenu dans la figure 4.1.

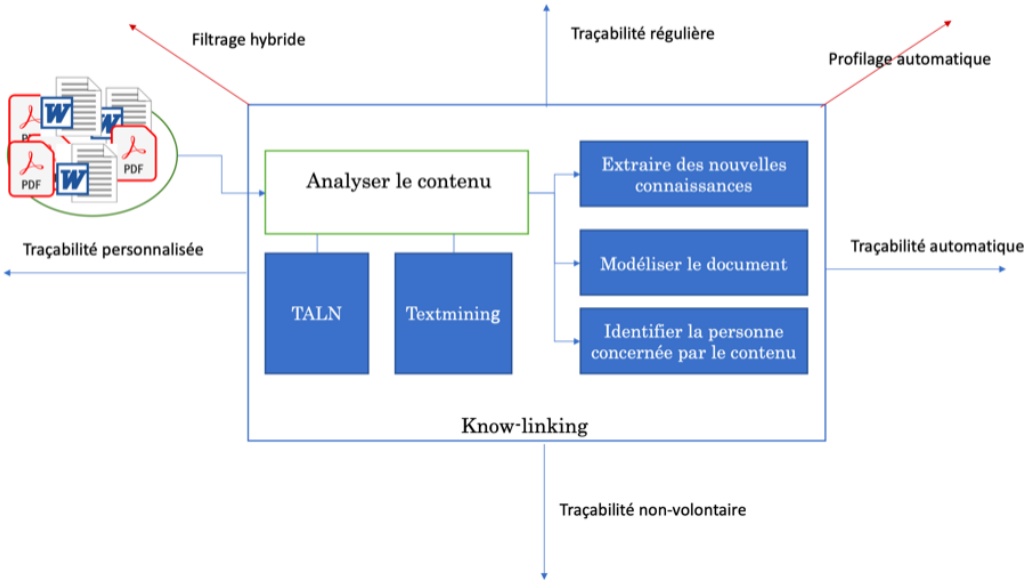


FIGURE 4.1 – Les dimensions de Know-linking

Troisième partie

Une nouvelle approche de génération de connaissances partagées : know-linking

Chapitre 5

Génération des profils

"Dès l'arrivée, le départ se profile." Ylipe

Sommaire

5.1 Introduction

À l'issue de l'étude bibliographique que nous avons menée, l'approche de partage de connaissances que nous proposons est devenue plus concrète avec l'ensemble des dimensions qui répondent à nos différentes questions de recherche. L'approche Know-linking prend la dimension du dynamisme de la traçabilité régulière et automatisée, la dimension de personnalisation grâce aux techniques de profilage hybride de collaborateurs et elle est dotée d'une autonomie d'extraction et d'enrichissement du capital de connaissances de l'entreprise grâce aux techniques de l'analyse du contenu.
La première étape de cette approche Know-linking consiste à construire d'une façon automatique et semi-supervisée des profils de collaborateurs. Dans ce chapitre, nous détaillons la phase de génération des profils de collaborateurs.

5.2 Profil de collaborateur

Chaque collaborateur dans une entreprise est caractérisé par des éléments qui le distinguent parmi d'autres. Cependant, on peut trouver un groupe de collaborateurs qui partage des critères communs ce qui nous permet d'exprimer un profil.
Les critères que nous prenons en compte dans notre approche de profilage sont :

— **Les missions :** désignent l'ensemble des tâches exercées par le collaborateur. Pour chaque tâche ou action réalisée le collaborateur a besoin de consulter certaines sources de connaissances. Il en produit donc un certain nombre et il laisse ainsi une certaine trace. Identifier la liste des missions c'est donc identifier ces traces à travers les documents produits ou consultés pour chaque tâche.

— **L'organisation :** chaque collaborateur dans l'entreprise est rattaché à un département ou à une équipe. Cette position dans l'organisation de l'entreprise détermine sa spécialité (informatique par exemple s'il est rattaché à une direction des systèmes d'information).

— **Les outils :** l'ensemble des outils est l'environnement logiciel relatif à un profil de collaborateurs. Cet environnement est un support à la production et au transfert de connaissances.

— **Liste de concepts :** l'ensemble des concepts du vocabulaire décrivant le profil.

Nous illustrons la structure globale d'un profil dans la figure 5.1 suivante.

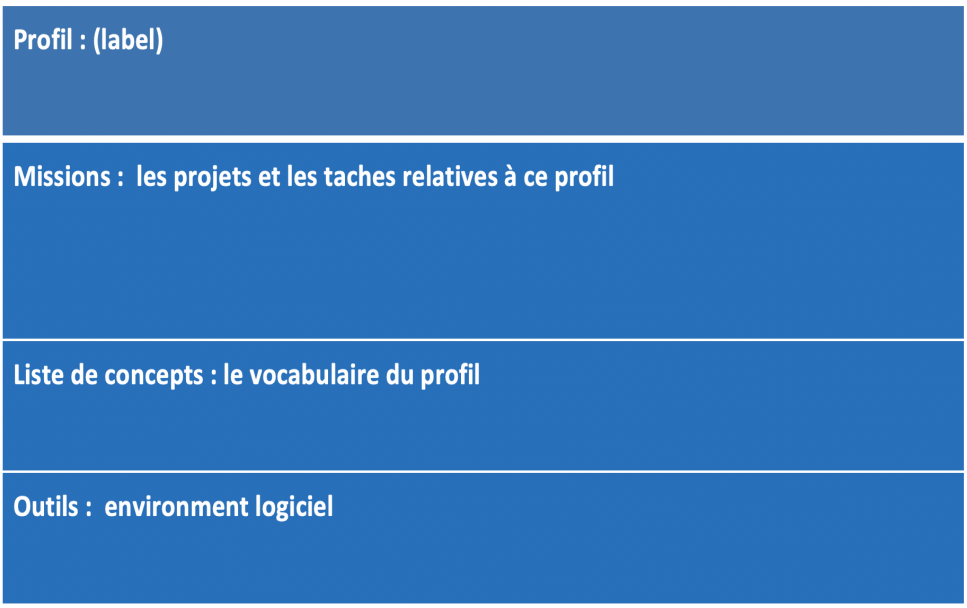


FIGURE 5.1 – La structure d'un profil

La structure du profil d'un collaborateur que nous proposons combine deux dimensions différentes :

— **La partie organisationnelle :** nous considérons la position du collaborateur dans l'organisation de l'entreprise, par exemple l'équipe et le département dans lesquels il évolue et

avec lesquels il partage ses connaissances. Ces informations sur la position des collaborateurs peuvent être accessibles dans les outils et les supports des ressources humaines.

— **La partie opérationnelle :** les projets dans lesquels il est impliqué, les activités et les opérations qu'il réalise dans le processus métier de l'entreprise. Ces informations sont enregistrées généralement dans les outils de travail de l'équipe, par exemple les outils de gestion du projet et la répartition des tâches.

Afin d'aboutir à la construction de cette structure de profils, deux analyses différentes doivent être effectuées : **une analyse organisationnelle** et **une analyse opérationnelle**.

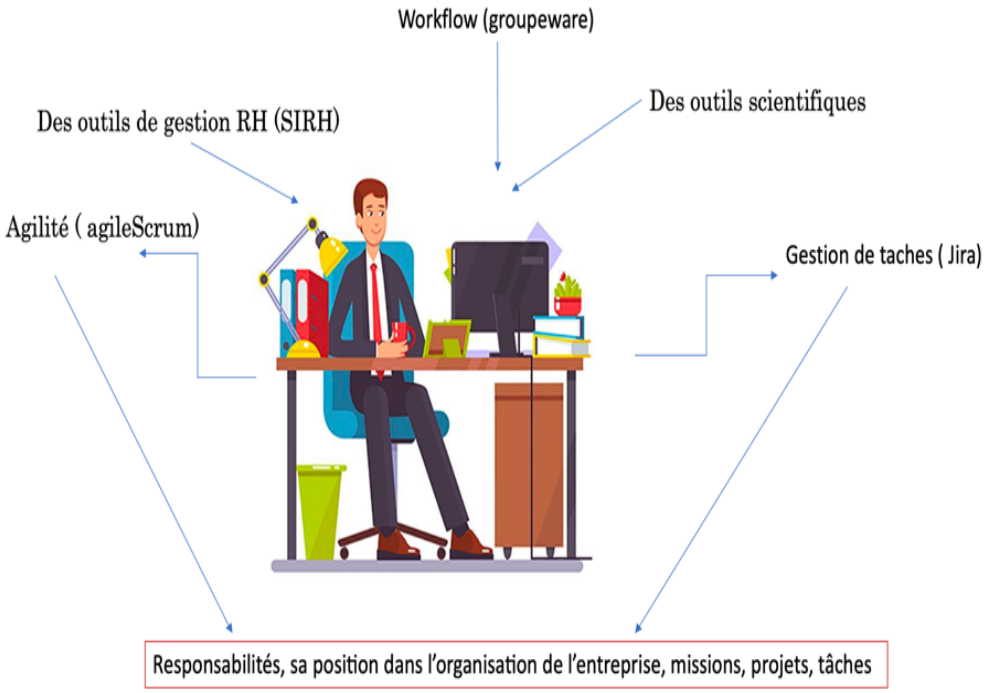


FIGURE 5.2 – L'environnement du travail d'un collaborateur

5.3 Analyse organisationnelle

L'organisation des ressources humaines d'une entreprise est un processus qui commence par le recrutement d'un employé jusqu'à son départ avec entre-temps sa formation et son évolution au sein de l'entreprise. Ce processus englobe les différentes positions qu'un employé peut occuper, les départements et les équipes qu'il a intégrés tout au long de son parcours.
Aujourd'hui il existe des systèmes et des outils de travail dédiés au stockage des informations relatives aux différentes démarches de la gestion et de l'organisation des ressources humaines. L'analyse de ces outils et des documents des ressources humaines permettent d'extraire la partie organisationnelle du profil du collaborateur, autrement dit de savoir « ce que le collaborateur est censé faire ». On peut classer ces supports de stockage et de gestion de ressources humaines en deux : supports logiciels et supports documentaires. Parmi ces supports nous remarquons que dans les entreprises il

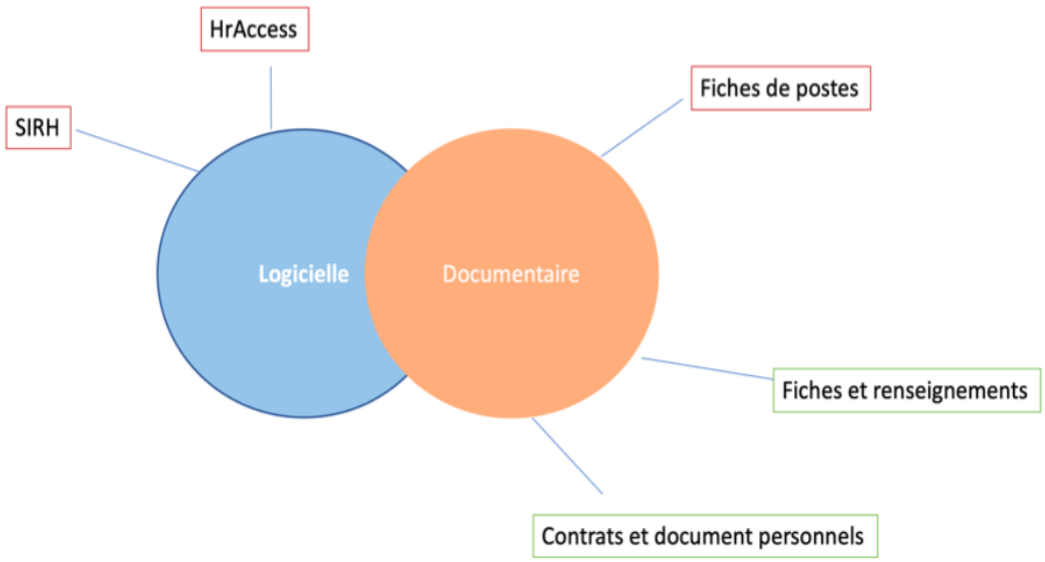


FIGURE 5.3 – Les supports de données ressources humaines

existe souvent des fiches de description de chaque poste. Ce sont des documents formels qui possèdent une structure standard.
La description de poste est un support qui décrit pour chaque poste l'ensemble des missions, son domaine, et qui donne d'autres informations comme les compétences requises, l'environnement techniques, etc. Les parties à analyser dans ces fiches sont illustrées dans la figure 5.4 ci-dessous.

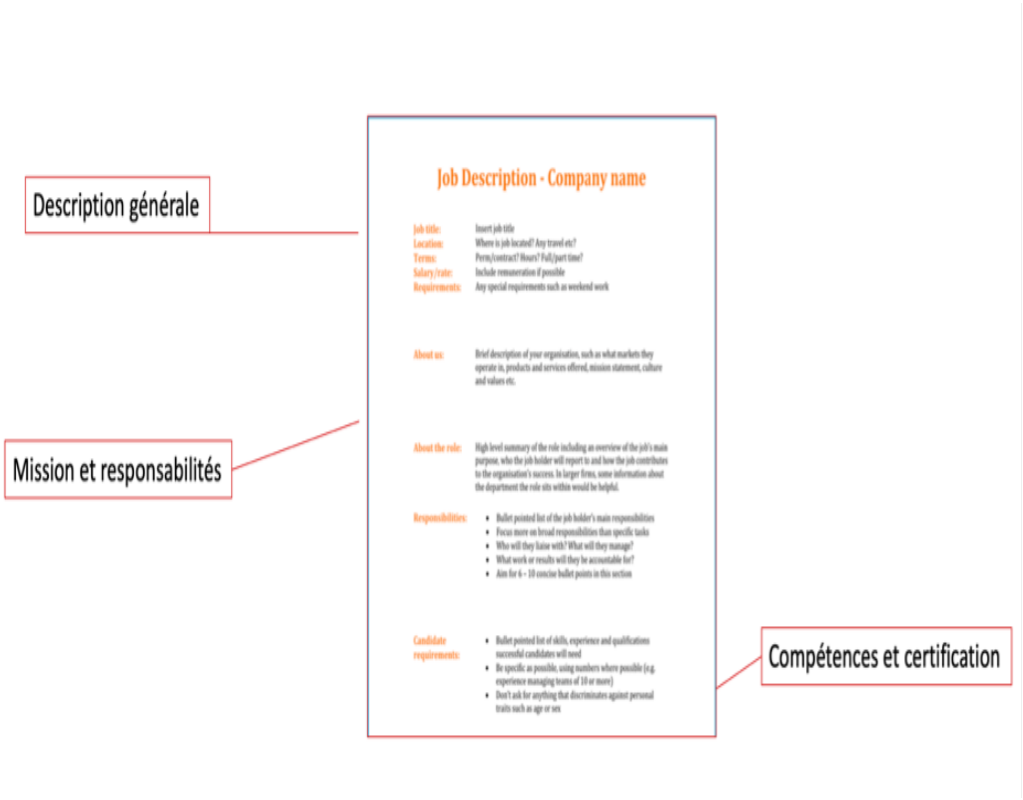


FIGURE 5.4 – La structure d'une fiche de description de poste

Ce type de document nous intéresse car il représente le point de départ pour la construction des profils de collaborateurs. Plusieurs travaux de recherche se sont basés sur la description des postes, ou en anglais « job description », dans le cadre du support des systèmes de recrutement électroniques «E-recruitment». Par exemple, [Ahmed Awan *et al.* 2019] propose d'analyser les job description afin de recommander au recruteur le meilleur profil du

candidat. L'analyse du contenu des descriptions de poste consiste à identifier les entités dans la description du poste qui sont généralement : Titre du poste, les compétences, les missions et le nombre d'années d'expérience [Ahmed Awan *et al.* 2019].
L'analyse de la description du poste peut se faire de trois manières :

1. **Techniques basées sur des règles et des modèles** [Ahmed Awan *et al.* 2019] : une approche basée sur des modèles prédéfinis pour analyser un texte non structuré. Par exemple, l'identification de l'adresse d'une personne peut nécessiter la présence de « habiter à » ou « résider à » dans le texte. Le problème de ces approches consiste en l'incapacité de découvrir de nouveaux modèles, ce qui signifie qu'en cas d'utilisation de nouveaux modèles non prédéfinis comme des synonymes ou de nouvelles phrases, des informations pourraient être perdues.
2. **Découverte de modèles à l'aide de techniques basées sur l'apprentissage automatique** [Bijalwan *et al.* 2014] : en se basant sur les modèles de Markov cachés (HMM) et les champs aléatoires conditionnels (CRF) [Ahmed Awan *et al.* 2019]. Ces approches nécessitent un grand nombre de données, sachant qu'elles ne parviennent pas à relier les informations avec le contexte.
3. **Techniques basées sur l'ontologie** [Vicient *et al.* 2011] : les approches basées sur l'ontologie utilisent des connaissances spécifiques au domaine pour extraire des informations significatives à partir de textes non structurés [Lafferty *et al.* 2001]. Si ces ontologies ne sont pas enrichies, les nouvelles informations extraites seront perdues. Pour un objectif différent nous utilisons la description de poste non pas pour rechercher le candidat approprié mais pour créer

des profils de collaborateurs. Pour analyser les descriptions de poste, nous proposons d'utiliser une méthode combinant les approches basées sur des ontologies et des modèles linguistiques proposée par [Ahmed Awan *et al.* 2019]. Ce qui représente une méthode automatique pour extraire des attributs textuels et les mapper dans des sources de connaissances structurées. Cette méthode nécessite une ontologie de domaine préliminaire. [Ahmed Awan *et al.* 2019] a présenté une ontologie pour modéliser la description de poste comme suit 5.5.

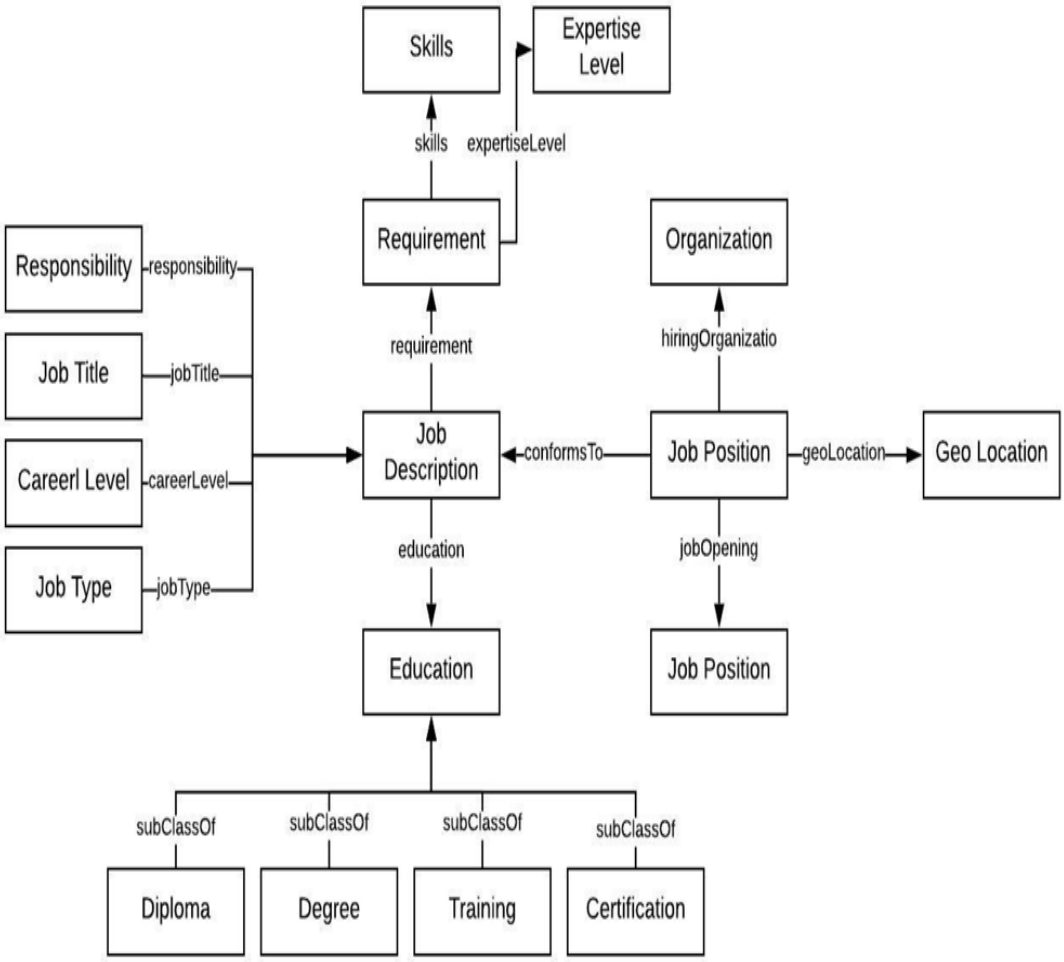


FIGURE 5.5 – Une Ontologie de description de poste [Ahmed Awan *et al.* 2019]

La deuxième entrée de cette approche est une liste de modèles (ou de règles). Nous en présentons quelques-unes dans le tableau 5.1

ci-dessous.

TABLE 5.1 – Exemple de règles.

Règles
Entité + poste ; /position=Directeur, gérant, assistant
«verbe à l'infinitif » + Mission
« Nom » + Mission
« Rattaché à » + Département
Spécialité + emploi ; / emploi = ingénieur, chercheur, technicien, designer

De nombreuses entreprises utilisent des outils spéciaux pour gérer leurs données ressources humaines telles que SIRH ou HR Access. Ces outils aident les entreprises à visualiser les données RH de chaque employé : son nom, âge, adresse, compétences, date de démarrage, certifications, service, vacances, arrêt maladie.
Les outils RH stockent les informations dans des bases de données avec quelques interfaces pour afficher et manipuler facilement certaines informations selon les besoins. Ainsi, afin d'extraire et d'analyser les informations RH, nous devons parcourir les bases de données liées. En parcourant et en analysant les bases de données du système, nous pouvons identifier pour chaque profil une liste d'informations supplémentaires. Analyser la description de poste et les outils RH ne suffit pas toujours pour construire une structure de profil complète car, dans la réalité, les experts sont impliqués dans de nombreux projets et nombreuses tâches qui n'ont pas été répertoriés précédemment dans les descriptions de poste. Les outils organisationnels représentent la théorie de l'organisation des experts, et les outils de gestion de projet représentent la partie pratique. Cela nous incite à analyser les outils de gestion de projet pour compléter les travaux précédents.

5.4 Identification du profil à partir des outils de gestion de projet de l'entreprise :

Les outils de gestion de projet sont utilisés dans les entreprises pour organiser la réalisation des projets. Cette notion de l'opération complète, celle de l'organisation où on peut savoir « ce que l'employé est censé faire » par « ce que l'employé est réellement en train de faire ». Chaque employé de l'entreprise est affecté à une liste de tâches définies dans les outils de gestion de projet. JIRA, agile scrum, kanban et autres sont utilisés dans ce contexte. Les outils mentionnés précédemment stockent leurs données dans des bases de données relationnelles, objets, des tableaux avec des lignes et des colonnes.
Nous proposons dans ce cas d'identifier une liste d'experts ainsi que leurs missions en analysant ces bases de données. À titre d'exemple, Laravel gitscrum est une application gratuite et open source pour les équipes agiles qui contient certaines fonctionnalités telles que le backlog du produit, le user story et le Sprint Backlog (figure 5.6).

Laravel giscrum stocke son contenu dans une base de données relationnelle. Pour analyser les informations manipulées par Laravel gitscrum, nous devrions procéder à l'analyse de la base de données de l'outil de gestion du projet. Cette analyse de la base de données qu'on propose utilise des requêtes de base de données comme SQL ou autre de type «SELECT» avec quelques jointures pour afficher les informations recherchées. Ces techniques d'extraction de données sur les profils sont exploitées dans un algorithme de profilage que nous détaillons ci-après.

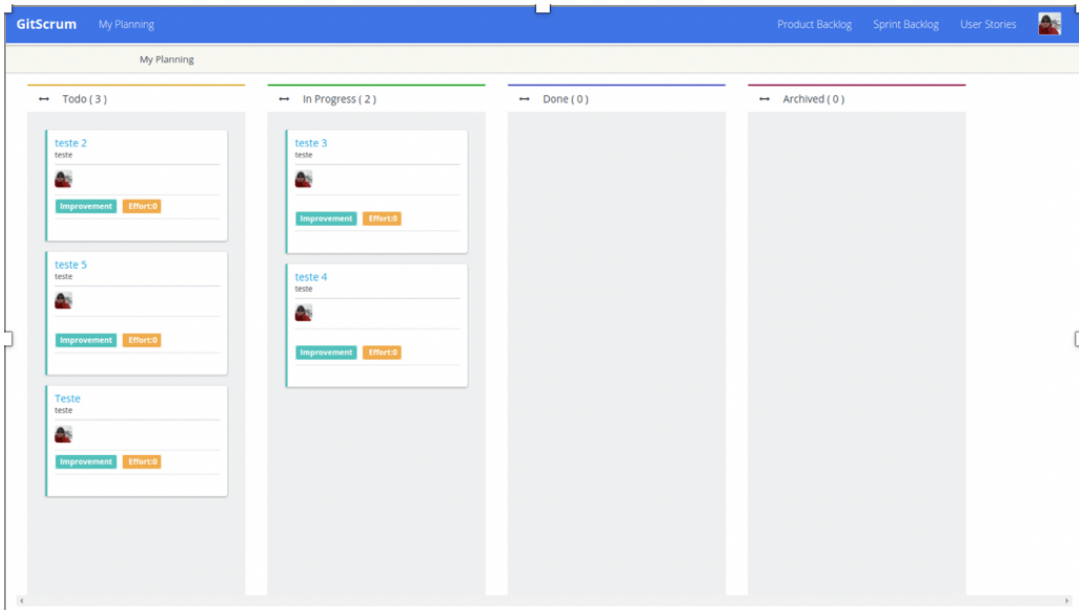


FIGURE 5.6 – Interface GitScrum

5.5 Algorithme de profilage

5.5.1 Les sources de données

Les données pour le profilage peuvent être collectées à partir de sources différentes : les bases de données relatives aux outils du travail, que ce soit opérationnel ou organisationnel, ainsi que les documents de ressources humaines comme nous l'avons précisé précédemment. En fonction de leur volume, ces données peuvent être extraites et stockées sous forme d'enregistrements ou de tuples dans des bases de données ou dans des fichiers de format excel ou CSV. Si le volume de ces données est beaucoup plus important, il faudra préparer une structure de type Data Lake [Miloslavskaya & Tolstoy 2016].

5.5.2 Les données d'entrée :

Dans le cadre du filtrage semi-supervisé que nous adoptons dans notre approche, nous commençons par définir une première liste de profils à partir de l'analyse des outils de l'organisation, et plus

particulièrement à analyser les job descriptions et les transformer en profils.
Le principe consiste à commencer par une liste de profils et à en ajouter d'autres en analysant une quantité de données.

La bibliothèque de Natural langage processing «NLTK» permet de nettoyer le texte de la job description ou de la « description de poste » puis d'identifier et d'extraire les entités linguistiques selon des règles comme présenté dans la partie précédente (analyse de fiches de poste). Nous ajoutons quelques règles à considérer dans cette phase d'analyse comme présenté dans le tableau 5.2 suivant.

TABLE 5.2 – Exemple de règles.

Règles
Si (terme = « Titre ») alors le successeur est « label » -> label
Si (terme = « Intitulé du poste ») alors le successeur est « label » -> label
Si (terme = « Titre du poste ») alors le successeur est « label » -> Label
Si (terme = « Responsabilités ») et Si (successeur = liste de verbes à l'infinitif) alors le suivant est liste de mission -> Liste de missions
Si (terme = « Responsabilités ») et Si (successeur = liste de noms) alors le suivant est liste de mission -> Liste de missions
Si (terme = « Missions ») et Si (successeur = liste de verbes à l'infinitif) alors le suivant est liste de mission -> Liste de missions
Si (terme = « Missions ») et Si (successeur = liste de Noms) alors le suivant est liste de mission -> Liste de missions
Si (terme = « Tâches ») et Si (successeur = liste de Noms) alors le suivant est liste de mission -> Liste de missions
Si (terme = « Tâches ») et Si (successeur = liste de verbes à l'infinitif) alors le suivant est liste de mission -> Liste de missions
Si (terme = «actions ») et Si (successeur = liste de verbes à l'infinitif) alors le suivant est liste de mission -> Liste de missions
Si (terme = «travail à faire») et Si (successeur = liste de verbes à l'infinitif) alors le suivant est liste de mission -> Liste de missions

Cette analyse nous permet d'obtenir les premières structures de profils et de passer de la description de poste à un profil de collaborateur comme illustré dans la figure 5.7.

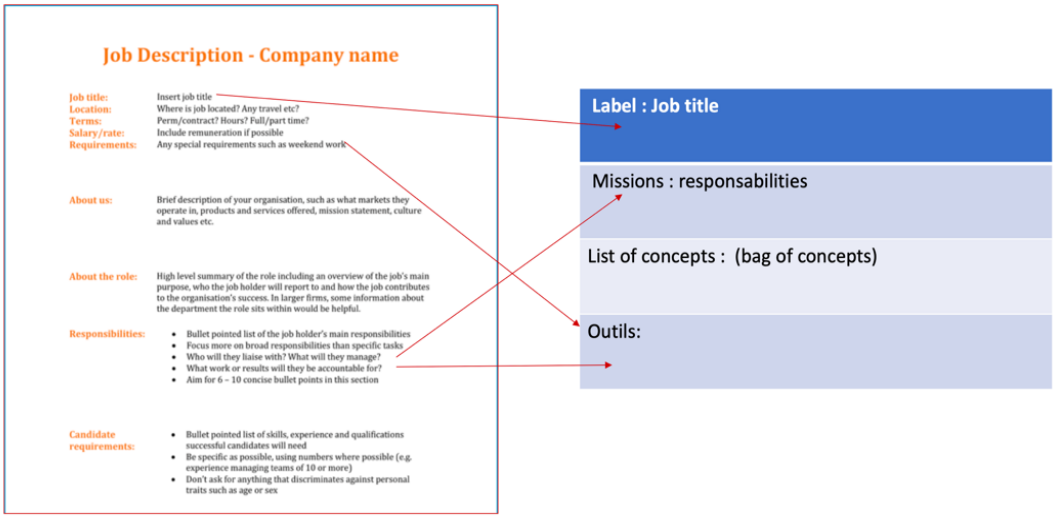


FIGURE 5.7 – Modèle de traduction d'une description de poste en profil de collaborateur

5.5.3 Principe de l'algorithme de profilage semi-supervisé :

L'objectif de l'algorithme de profilage est de construire autant de structures de profils que possible en se basant sur quelques structures prédéfinies. Les instructions de l'algorithme sont les suivantes :

1. Parcourir la liste de profils initiaux élément par élément ;
2. Comparer les données d'un enregistrement I de données collectées (X) avec un élément S de la liste de profils ;
3. Vérifier la similarité entre les deux éléments I et S ;
4. Si les éléments sont considérés similaires alors I est un acteur de profil S. Dans ce cas on ajoute le label de S devant l'enregistrement I ;
5. Si les éléments ne sont pas similaires alors on obtient un nouveau profil et on l'ajoute à la liste des profils.

TABLE 5.3 – Variables d'entrée

Variable d'entrée	Description
X	L'ensemble de données des collaborateurs collectées sous forme des enregistrements .
Liste_de_profils	La liste de quelques profils préparée au préalable comme jeu de données .
I	Un enregistrement composé des champs : mission, label, liste de concepts et domaine.
S	Un enregistrement composé des champs : missions, label, liste de concepts et domaine.

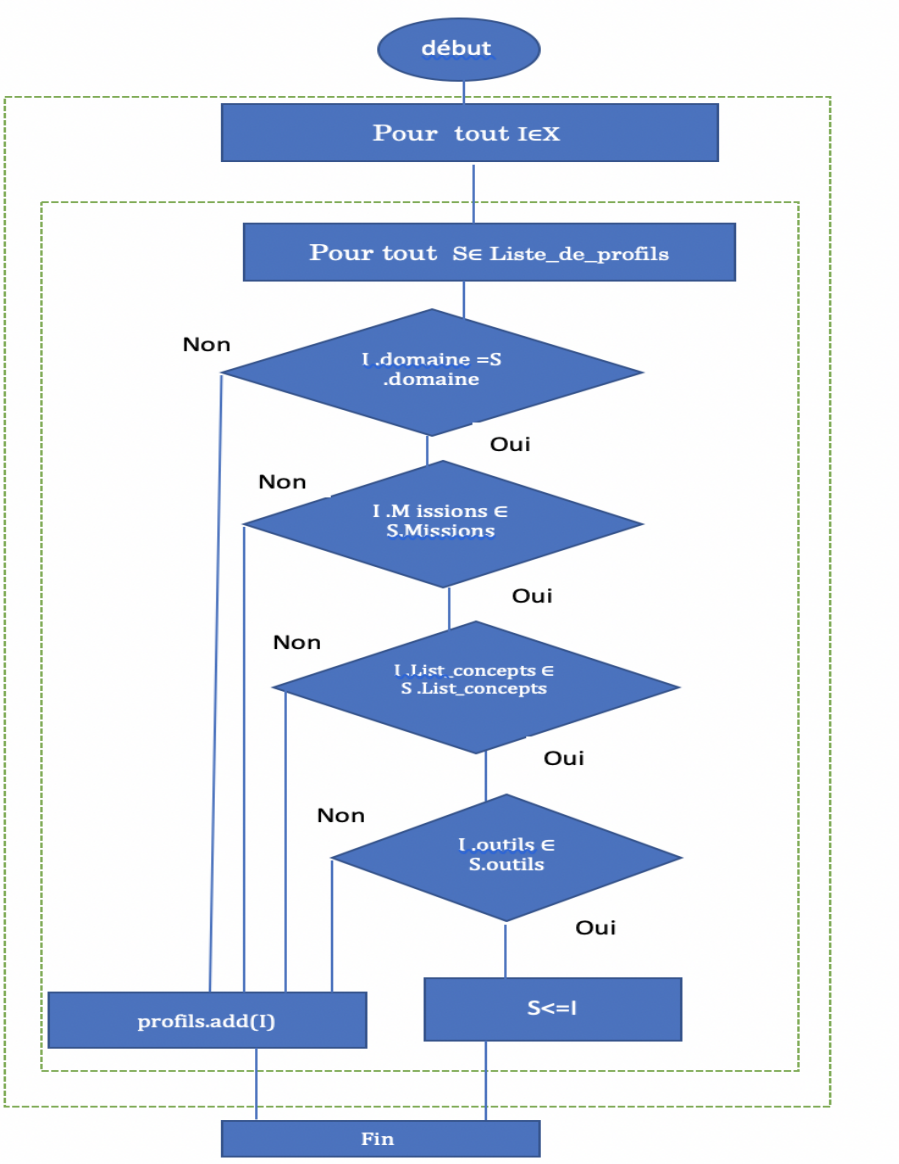


FIGURE 5.8 – Algorithme de profilage

5.5.4 Calcul de similarité :

La fonction que nous avons utilisée pour le calcul de similarité entre les missions de «I» et les missions de «S» se fait à l'aide des dictionnaires linguistiques tels que WordNet et un glossaire de l'entreprise, ce qui permet de trouver les synonymes et les hyponymes. Le glossaire de l'entreprise permet d'enrichir ce calcul de similarité tout en considérant le vocabulaire propre à l'entreprise. Par exemple, à l'aide de ce glossaire, on peut détecter que «DSI» et «département des systèmes d'informations» sont synonymes, ou encore que «KM» est utilisé pour dire manager des connaissances (knowledge manager).Nous proposons alors dans le schéma de la figure 5.9 un algorithme de similarité qui se base sur ce principe.

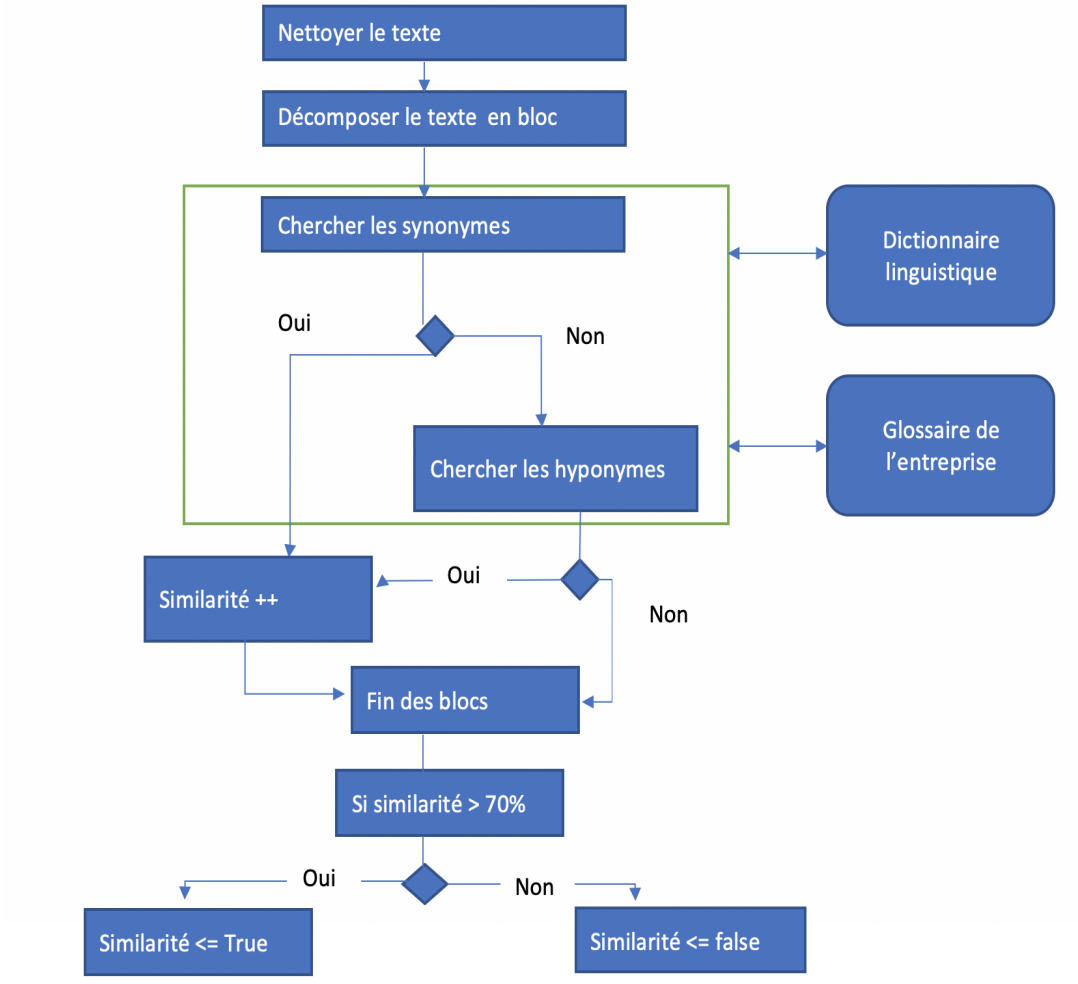


FIGURE 5.9 – Algorithme de similarité

5.6 Conclusion

L'analyse du contenu textuel est utilisée comme principe de base pour élaborer dans notre algorithme de profilage semi-supervisé des profils de collaborateurs précis et concrets tout en considérant les dimensions organisationnelles et opérationnelles des collaborateurs.

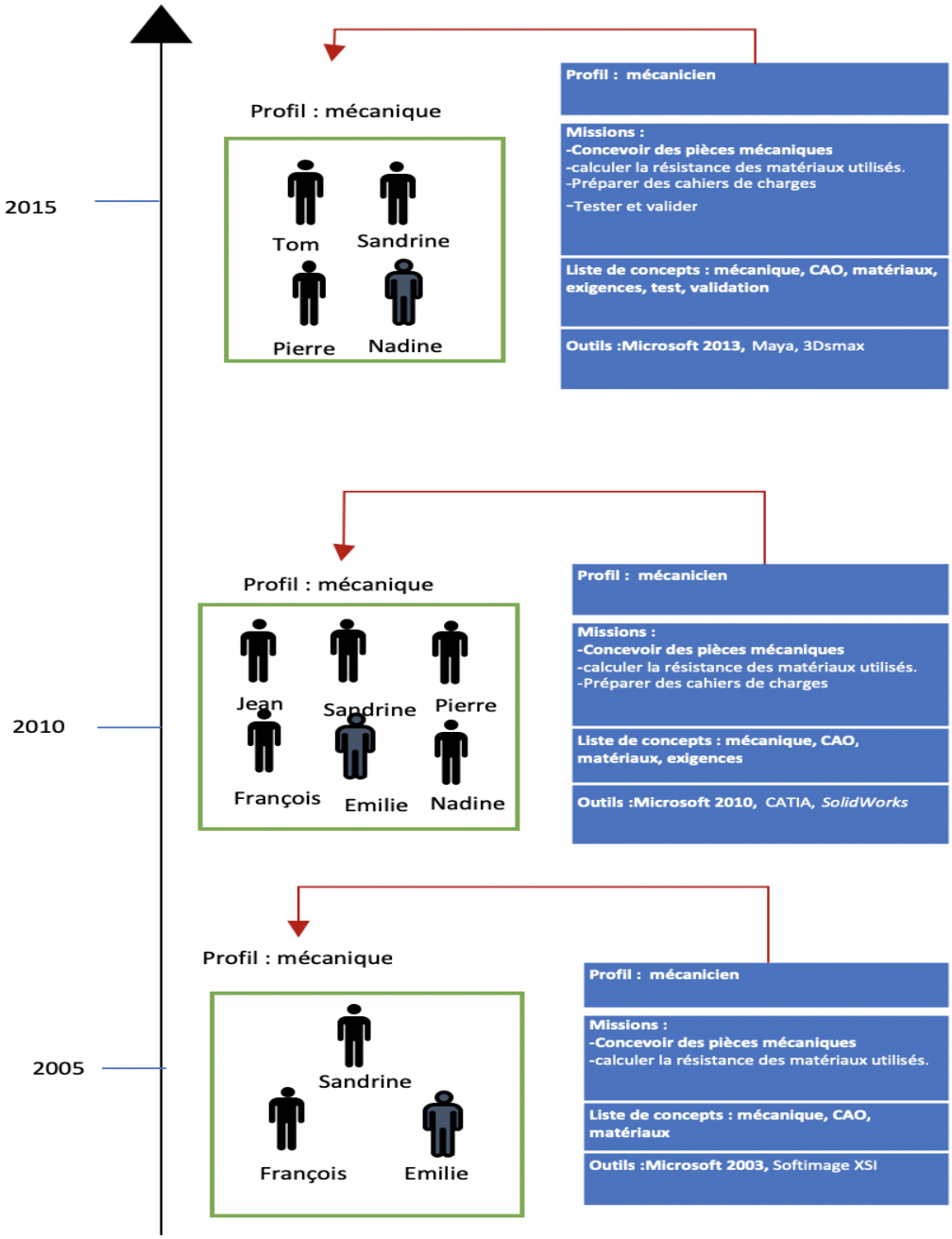


FIGURE 5.10 – Évolution et traçabilité

Dans la figure 5.10 nous illustrons l'exemple d'un profil (mécanique) dans une entreprise à des périodes différentes à savoir en 2005, en 2010 et en 2015. Cette capture du même profil de collaborateurs à trois reprises met en évidence que les missions, les outils et les concepts qui décrivent un profil ont évolué dans le temps parallèlement à l'évolution organisationnelle de l'entreprise. En effet, on peut remarquer que certaines personnes ont quitté l'entreprise et que d'autres l'ont intégrée. Notre approche de génération de profils permet de capter automatiquement cette évolution. En analysant cette capture (figure 5.10) du même profil on peut noter que la génération de profil est un garant de plusieurs points dans notre approche :

- La traçabilité personnalisée : En suivant l'évolution des profils on peut garder les traces des collaborateurs dans le temps et d'une façon personnalisée (comme illustré dans la figure 5.10).
- L'exploitation de l'expérience passée : La distance temporelle intégrée à la structure du profil assure au collaborateur l'utilisation de l'expérience passée que son collègue a produite il y a des années.
- La compréhension des besoins des collaborateurs en tant que source de connaissances : la description du profil par les missions, les concepts et les outils simplifie la compréhension des besoins en documents. Cette notion a fait l'objet d'un chapitre ultérieur de ce manuscrit sur la distribution et la classification de documents.

Dimensions à prendre en compte :

Les dimensions que notre approche doit prendre en compte dans sa prochaine étape sont l'ensemble des échanges (en termes de connaissances) entre les profils.

Ces échanges peuvent être tracés à travers deux types de lien entre

les profils :

— Un lien dans le temps : Le lien dans le temps s'effectue lorsque un collaborateur de profil X cherche une connaissance produite il y a quelques années par un autre collaborateur du même profil.

— Un lien dans l'espace : quand un collaborateur de profil X cherche une connaissance produite par un collaborateur de profil Y et que les deux collaborateurs travaillent dans l'entreprise durant la même période, et ce généralement dans le cadre de collaborations entre plusieurs spécialités pour réaliser un projet.

L'exploration de ces liens complète le travail de profilage par l'ensemble des échanges entre les profils ce qui engendre la génération de nouvelles connaissances partagées.

Chapitre 6

Génération des graphes de profils et des liens sémantiques pour la distribution des documents

"Comment vous représentez-vous ça : l'endroit où l'espace se termine ?" Arno Schmidt

Sommaire

6.1 Introduction

Retrouver la bonne connaissance dans un environnement complexe et hétérogène est un sujet classique qui se compose de deux parties : la première consiste à trouver une bonne stratégie d'organisation des documents sources à l'aide des techniques d'indexation ou de classification. La deuxième est de définir une méthode efficace de recherche de ce document (par exemple pouvoir reformuler des requêtes pertinentes). Par conséquent, la résolution du problème d'extraction d'informations « information retrieval » est liée principalement à la bonne organisation des documents et à la bonne requête de recherche.

FIGURE 6.1 – Modèle classique de recherche d'information

Cette vision classique du problème (dans la figure 8.18), nous a poussés à approfondir ce problème afin de trouver une réponse à la question de recherche que nous avons déjà posée « comment réduire le bruit dans la recherche d'informations ? » Ainsi, une autre vision de la problématique que nous proposons dans notre approche consiste à résoudre le problème dans le sens inverse. En effet, au lieu de rechercher les connaissances par le collaborateur, ce sont les connaissances qui vont être recommandées à ce dernier afin de s'intégrer dans son environnement de travail, ce qui est communément appelé « Push ». Tout d'abord, nous définissons l'accessibilité des supports de connaissances comme étant un lien qui doit être pertinent entre le collaborateur et le document. La

logique de résolution du problème de la recherche de connaissances que nous proposons exige trois étapes :

— Comprendre le besoin des collaborateurs en documents (supports de connaissances) et définir des structures (un travail de profilage réalisé dans le chapitre génération des profils). Ces structures permettent d'élaborer une modélisation des besoins d'une façon compréhensible.

— Rechercher les liens entre les profils qui sont exprimés d'une manière implicite dans le contenu textuel.

— Respecter le besoin des collaborateurs et leurs interactions pour distribuer les documents

Dans ce chapitre, nous détaillons la deuxième et la troisième phase de notre approche consistant à générer d'une façon dynamique une modélisation des profils sous forme de graphes de connaissances, à analyser le contenu pour la découverte et la génération des liens sémantiques entre les profils de collaborateurs, et enfin à respecter cette structure pour le partage des documents.

6.2 Génération des graphes des connaissances

6.2.1 Génération des graphes des profils

6.2.1.1 Identifier le lexique du profil

Chaque profil de collaborateur se distingue par un vocabulaire spécifique qui consiste en un ensemble de termes en rapport avec son métier ainsi que son environnement technique. Les méthodes de l'analyse du contenu, et notamment la méthode bag of words [Kim *et al.* 2017], décrit le texte sous forme d'un ensemble de termes les plus fréquents. Cependant cette méthode ne prend pas en considération le traitement de la synonymie ou de l'hyperonymie. Par exemple, dans la fig.2, on a un ensemble de mots qui

décrivent un texte par la technique « bag of words » avec leur fréquence d'apparition.

TABLE 6.1 – Résultats de la méthode bag of words

Chat	Automobile	Oiseau	Animal	Véhicule
3	5	3	4	4

On remarque qu'il existe des relations sémantiques entre les termes présentés dans le tableau. Ces relations sont de types synonymie ou hyperonymie. Nous soulignons ces relations dans la figure 6.2 suivante.

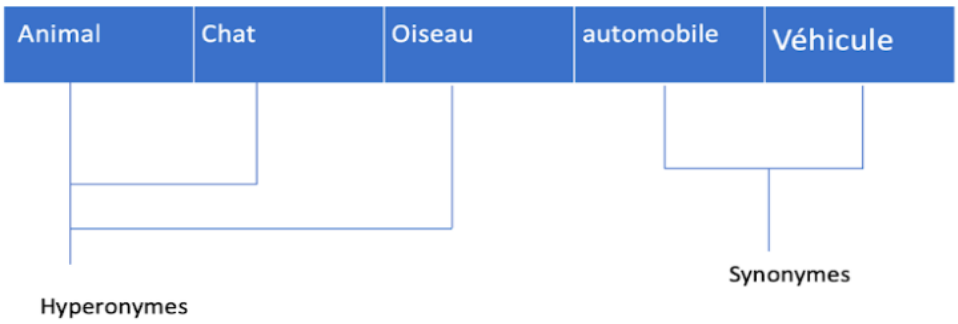


FIGURE 6.2 – Relation entre les termes

Le sac de concepts, ou bag of concepts [Kim *et al.* 2017], permet de résoudre le problème de cette représentation du texte en considérant les relations sémantiques qui peuvent exister entre les termes qui sont désormais appelés « des concepts ».
Cette méthode de représentation est utile pour représenter les profils de collaborateurs ainsi que les documents sous forme d'un «bag of concepts».
Le bag of concept [Kim *et al.* 2017] est une évolution de la méthode «bag of words» dans le sens où les liens entre les termes sont pris en considération. La méthode «bag of words» consiste à trouver les termes les plus fréquents dans un corpus. Elle résulte en un nombre de termes fréquents suivi de leurs nombres d'occurrences dans

un corpus donné. Les dimensions des tableaux générés peuvent être alors très larges, une raison qui a motivé les chercheurs pour les améliorer en travaillant les liens qui existent entre les termes comme la synonyme et l'hyperonyme. Le bag of concepts regroupe les termes similaires dans un même concept [Kim *et al.* 2017]. En se basant sur une ressource linguistique comme WordNet, les liens de similarité seront identifiés. Par conséquent la représentation d'un document est désormais la fréquence d'apparition de ces concepts [Kim *et al.* 2017].

Dans notre approche, nous considérons que pour tout profil X il existe une description E relative à ce profil telle que chaque élément i de l'ensemble E est un concept extrait à partir des documents structurés. Les documents structurés de ressources humaines forment la source sur laquelle nous nous basons afin d'appliquer la méthode bag of concepts. En effet, ces concepts, en lien avec la mission et les activités opérationnelles du collaborateur, peuvent représenter les premiers éléments du glossaire du domaine du profil E (figure 6.3 ci-dessous).

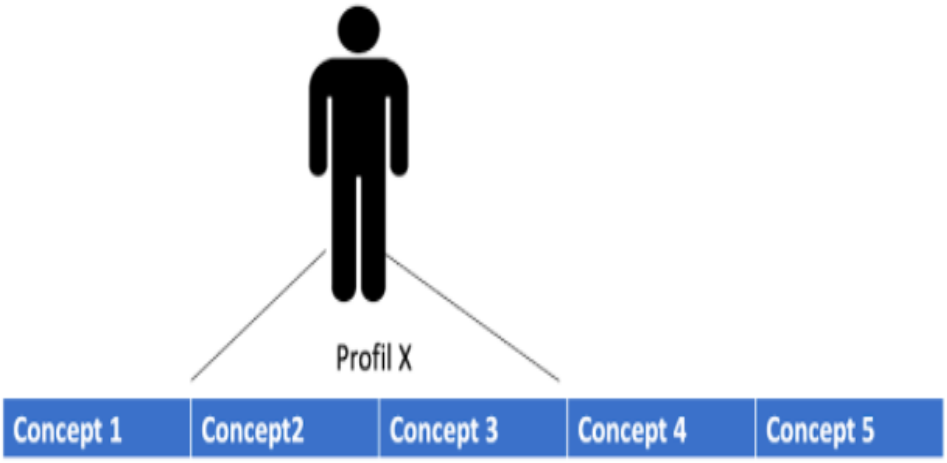


FIGURE 6.3 – Bag of concepts d'un profil

Nous illustrons le processus d'identification du lexique du profil dans la figure 6.4 suivante.

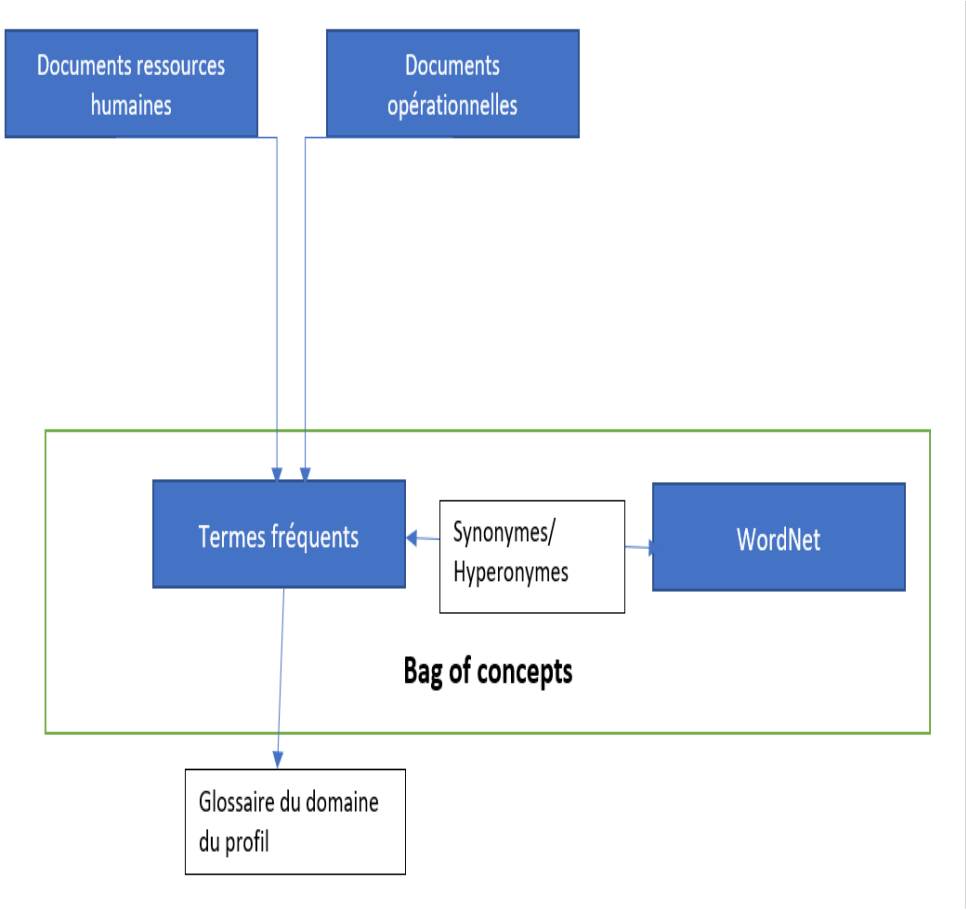


FIGURE 6.4 – Processus d'identification du lexique des profils

6.2.1.2 Déterminer les liens entre les concepts du graphe

Les techniques de l'analyse des langages naturels permettent de distinguer les verbes et les noms dans une phrase, plus précisément le POS-tagging permet de distinguer les verbes et les noms dans une phrase à travers une labélisation suite à une analyse morpho-syntaxique. Les règles de passage du texte au graphe proposent de former les nœuds à partir des noms, et les verbes constituent les arcs, autrement dit les relations qui existaient entre les termes, de la façon suivante :

— Supposons que nous avons une phrase comme suit :

TABLE 6.2 – Structure d'une phrase

Terme 1	Verbe	Terme 2

— La transformation de cette phrase en graphe donne (figure 6.5).

FIGURE 6.5 – Une phrase sous forme d'un graphe

Cependant, cette génération fournit un graphe de taille très large et de faible performance. Il est donc important de filtrer les termes et de ne considérer que les concepts de profils, autrement dit de vérifier que le graphe ne contient que des concepts.
L'objectif de cette étape est de générer une représentation pour chaque profil sous forme d'un graphe de connaissances ne prenant en compte que les concepts. Pour ce faire, nous nous appuyons sur le glossaire du domaine du profil que nous avons préparé lors de l'étape précédente, précisément nous vérifions si le terme appartient à la liste des concepts générés par l'application de la méthode bag of concepts ou non. Si c'est le cas, il sera modélisé comme un nœud lié sémantiquement aux autres concepts. Nous illustrons ce passage du profil au graphe dans la figure 6.6 ci-dessous.
Les bibliothèques du Traitement de Langage Naturel ainsi que les outils de Textmining permettent d'extraire les termes d'un texte, puis un graphe de termes peut se construire en liant le lexique du profil (l'ensemble des termes), qui sera affiné afin de tracer le graphe de concepts liés à chaque profil. Au final, pour chaque Profil un modèle de représentation sera complété comme indiqué dans la figure 6.7.

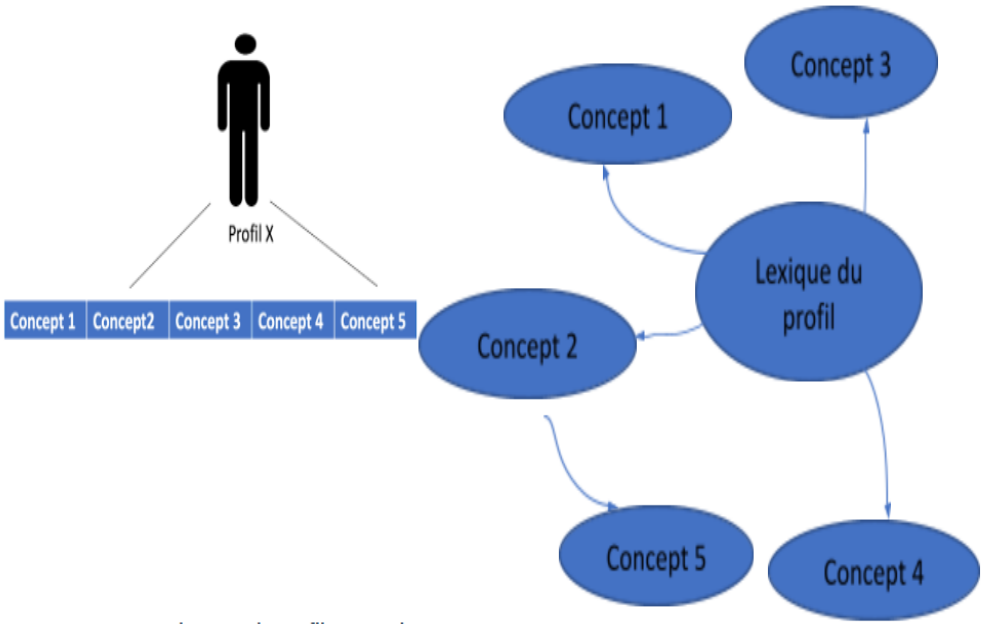


FIGURE 6.6 – Du profil au graphe

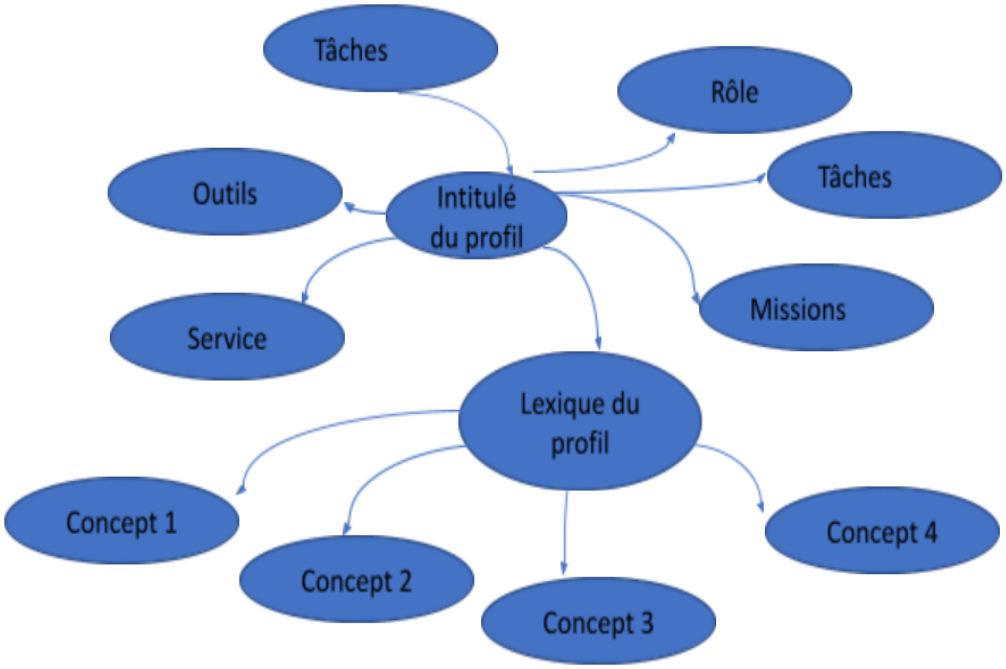


FIGURE 6.7 – Structure du graphe de profil

6.2.2 Le graphe de connaissance de l'entreprise

En combinant les graphes des profils ensemble nous obtenons un graphe de connaissance de l'entreprise. Ce graphe est une modélisation semi-formelle des connaissances de l'entreprise considérant

les différents profils des collaborateurs ainsi que les concepts des domaines sur lesquels ils/elles travaillent.
Nous illustrons un exemple de graphe de connaissance de l'entreprise dans la figure 6.8 ci-dessous.

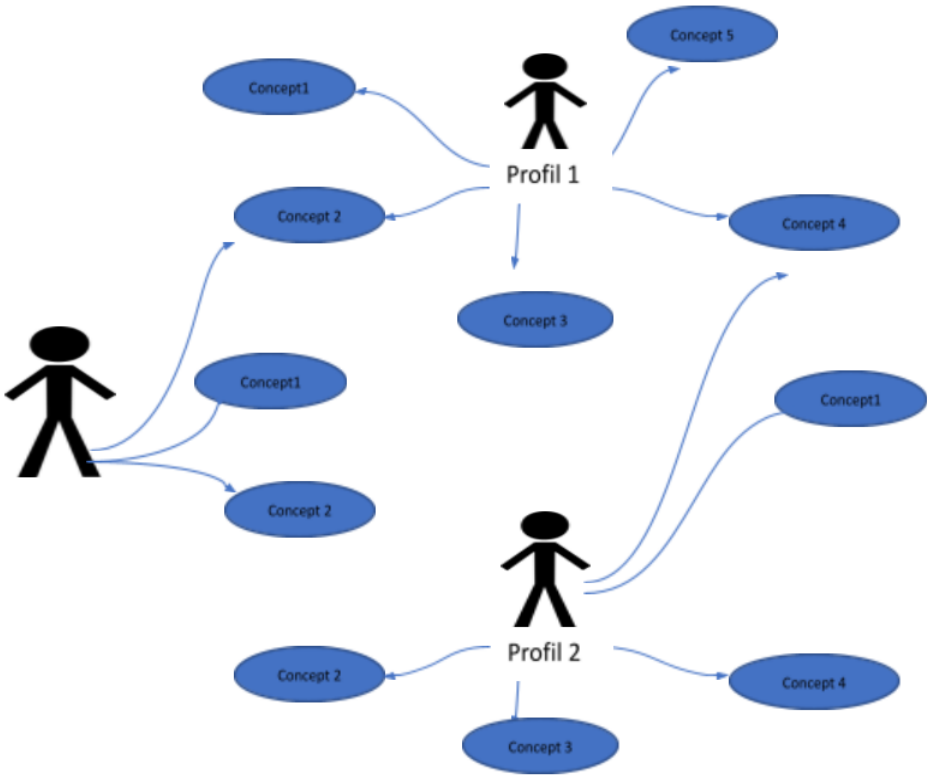


FIGURE 6.8 – Graphe de connaissances de l'entreprise

6.2.3 Enrichissement automatique des graphes des connaissances

La liste des concepts générés initialement est un ensemble de concepts générés au moment de l'application de l'algorithme de profilage. Cependant, le processus de production de documents continue à être exécuté d'une façon continue, ce qui engendre au fur et à mesure de nouveaux concepts. Un «gap» se construit petit à petit entre les concepts générés et les nouveaux concepts exprimés dans les nouveaux documents. Dans notre approche, la liste

des concepts relative à un profil est enrichie en permanence et dynamiquement par l'application de la méthode de bag of concepts sur les nouveaux documents produits par les profils au fur et à mesure, ce qui permet de mettre à jour ces profils par de nouveaux concepts. Nous modélisons cette étape dans la figure 6.9 suivante :

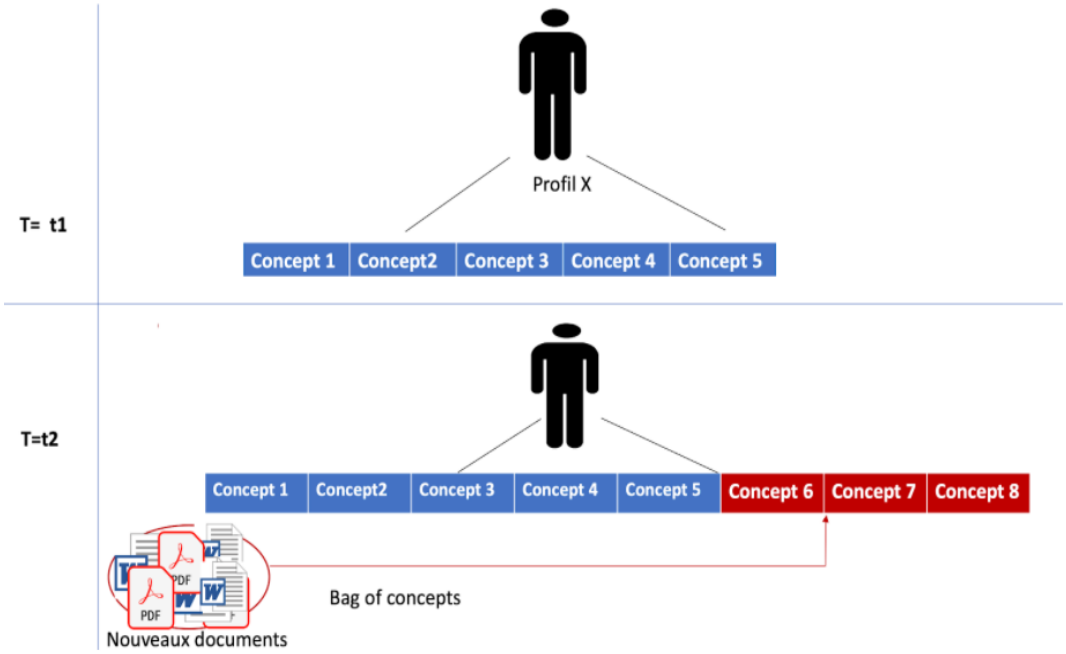


FIGURE 6.9 – Bag of concepts enrichi

Dans la figure 6.9 nous captons l'évolution des concepts d'un même profil à deux moments différents tels que la variable temps T= t1. Nous avons les concepts= concept1, concept2, concept3, concept4, concept5. Après un certain temps et une deuxième capture à temps T=t2, nous remarquons que la liste a évolué et que de nouveaux concepts sont apparus tels que T=t2 et les concepts= concept1, concept2, concept3, concept4, concept5, concept6, concept7, concept8.

Nous illustrons toutes les étapes de l'enrichissement d'un graphe de connaissances dans la figure 6.10.

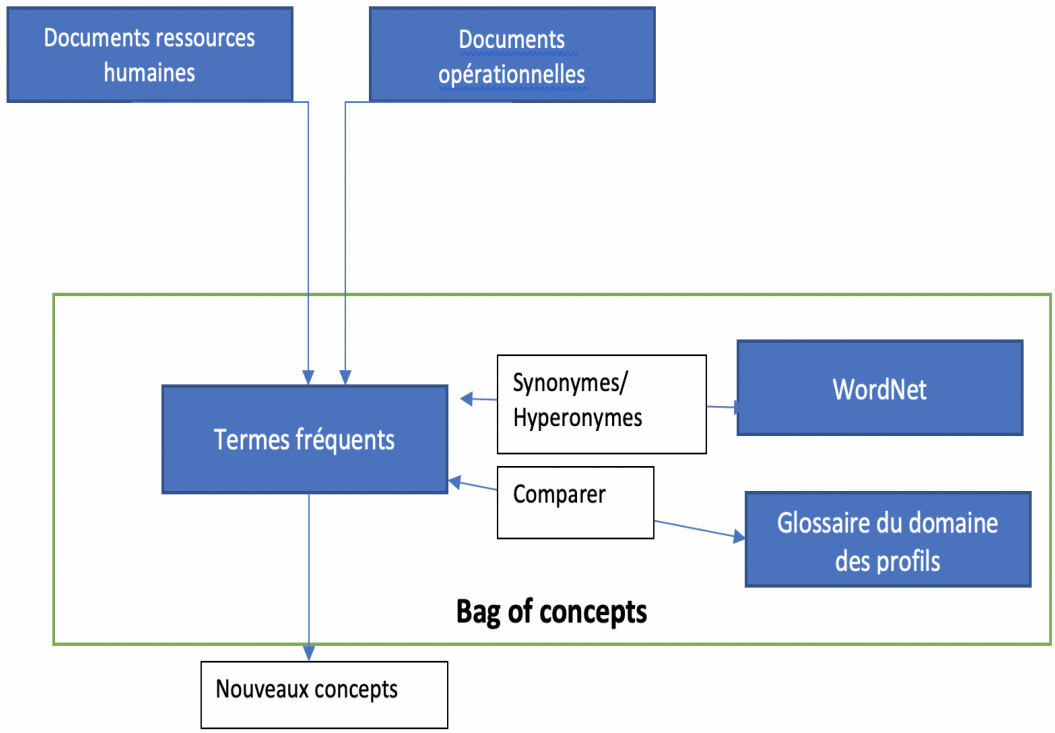


FIGURE 6.10 – Processus d'enrichissement des graphes de connaissances

6.3 Génération des liens sémantiques

6.3.1 Importance des liens sémantiques dans le partage de connaissances

Chaque profil de collaborateur est défini par un ensemble de concepts qui évoluent au fil du temps. Ces concepts sont liés dans le graphe représentatif des profils. Dans plusieurs cas, le collaborateur a besoin de consulter des supports de connaissances dans d'autres domaines, des supports que la définition du profil ne prend pas en compte. L'objectif de notre approche de génération des profils est de comprendre le besoin des collaborateurs en connaissances et de délimiter le cercle des supports de connaissances qui les concerne. Cependant, ce cercle couvre aussi des supports de connaissances produits par un autre profil. Nous allons expliquer de manière plus approfondie l'importance de la génération des liens

sémantiques entre les profils dans le partage de connaissances en suivant l'étude de l'exemple suivant.
Considérant que nous avons ce document (figure 6.11) produit par un chef de projet dans une entreprise et qui décrit le choix du matériel : si nous nous posons la question : «qui a besoin de la

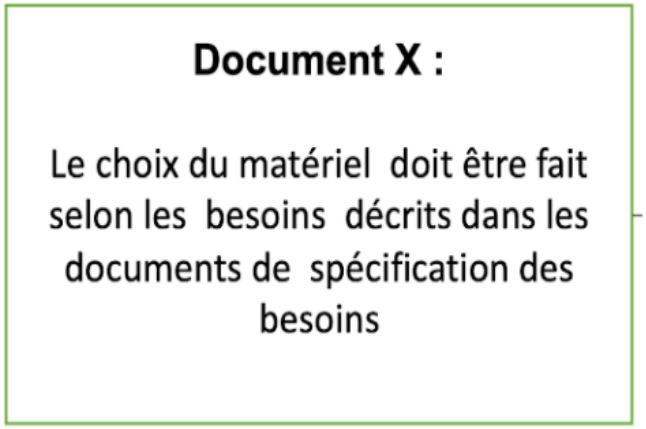


FIGURE 6.11 – Exemple d'un document contenant un lien sémantique

connaissance exprimée dans le document X ?», la réponse est la personne responsable des achats des matériaux. C'est bien elle qui devra choisir le nouveau matériel à acheter selon les spécifications décrites dans le document concerné. Par conséquent, elle doit avoir accès à ce document à l'instar du chef de projet (le producteur du document). L'application des algorithmes d'indexation classiques n'a pas pu détecter le besoin du commercial de l'entreprise de ce document. Ce dernier doit fouiller tous les documents des autres domaines pour trouver ceux qui le concernent.
Nous proposons donc de fouiller les liens sémantiques exprimés dans le document X. Nous pouvons ainsi constater que le bloc de texte comporte deux concepts : « choix du matériel », un concept qui appartient au profil de commercial, et spécification des besoins, un autre concept appartenant au profil chef de projet. Les deux

concepts sont liés dans un même bloc de texte, un lien sémantique est donc détecté.

6.3.2 Algorithme de génération des liens sémantiques

Afin de pouvoir identifier ces liens sémantiques, nous proposons un algorithme spécifique qui permet d'analyser un bloc de texte et de capter automatiquement les liens sémantiques qui existent et qui prend en entrée une carte de connaissances construite en combinant les graphes des profils.

6.3.2.1 Principe :

1. Décomposer le texte en phrases ;
2. Distinguer les verbes des noms de la même phrase ;
3. Prendre un premier terme c et vérifier s'il appartient aux concepts définis dans la carte de connaissances ;
4. Si c'est le cas, on identifie le profil correspondant ;
5. Extraire à chaque fois le profil d'un élément de la liste des mots, vérifier s'il est concept ou non et le comparer à c ;
6. Si on trouve deux concepts qui appartiennent à deux profils différents on sauvegarde le lien sémantique ;
7. Extraire un second mot et refaire les mêmes étapes (3,4,5,6) jusqu'à ce qu'il n'y ait plus de mots à comparer dans la liste des mots ;
8. Refaire les étapes 1, 2, 3, 4, 5, 6, 7 jusqu'à ce qu'il ne reste plus de phrases dans le document
9. Refaire toutes les étapes précédentes jusqu'à ce qu'il n'y ait plus de documents dans le corpus

TABLE 6.3 – Liste des données d'entrée

Variable	Description
Liste_des_mots	La liste de tous les mots dans une phrase du texte d'un document à analyser
Liste_des$_{D}ocuments$	Un corpus de documents à traiter

6.3.2.2 Données d'entrée

6.3.2.3 Les données de sortie de l'algorithme

TABLE 6.4 – Les données de sortie

Variable	Description
Liste_des_liens	La liste des liens sémantiques trouvés

6.3.2.4 Variables locales

TABLE 6.5 – Variables locales

X	Un document textuel
Phrase	Une phrase du texte
Phrase	Une phrase du texte
C, C1	Des concepts
Profil_c	Le profil auquel appartient le concept C
Profil_c1	Le profil auquel appartient le concept c1

6.3.2.5 Les instructions de l'algorithme de recherche des liens sémantiques

Les instructions de l'algorithme de recherche des liens sémantiques dans les documents prennent en entrée l'ensemble des variables détaillées dans le tableau 6.3 ci-dessus et retournent les

variables présentées dans le tableau 6.4. Les instructions de l'agorithme sont modélisées dans le diagramme dans la figure 6.12 suivante.

6.4 La distribution des documents

6.4.1 La distribution des documents dans la littérature scientifique

Les documents restent l'une des sources les plus importantes de connaissances. Certaines techniques basées sur le clustering [Roul 2018], proposent une extraction des informations pertinentes à partir d'une liste de documents identifiés par un moteur de recherche. Ces approches consistent à identifier pour chaque document une liste d'attributs (mots contenus et leur poids), après une phase de nettoyage afin de sélectionner le cluster sémantique le plus proche en comparant les attributs des documents.
[Wang & Koopman 2017] ont proposé de rechercher des similitudes entre des articles scientifiques dans les domaines de l'astronomie et de l'astrophysique en construisant une structure sémantique basée sur les éléments définissant un article (auteur, domaine, titre et références). Les articles sont représentés dans un espace de mots où chaque article est défini par ses mots et leur poids. En effet, ils ont proposé de représenter pour chaque entité un vecteur de poids des termes, sachant qu'une entité peut faire référence à un auteur, une revue, un titre ou une référence. Afin de construire un groupe de documents, les auteurs ont proposé d'appliquer deux algorithmes différents : K-means et Louvain. Une phase de validation est proposée par intervention d'experts ou de références de groupe pour chaque groupe. Deux documents similaires ayant le même contenu mais représentés avec des mots différents ne sont pas reconnus en utilisant ces approches.

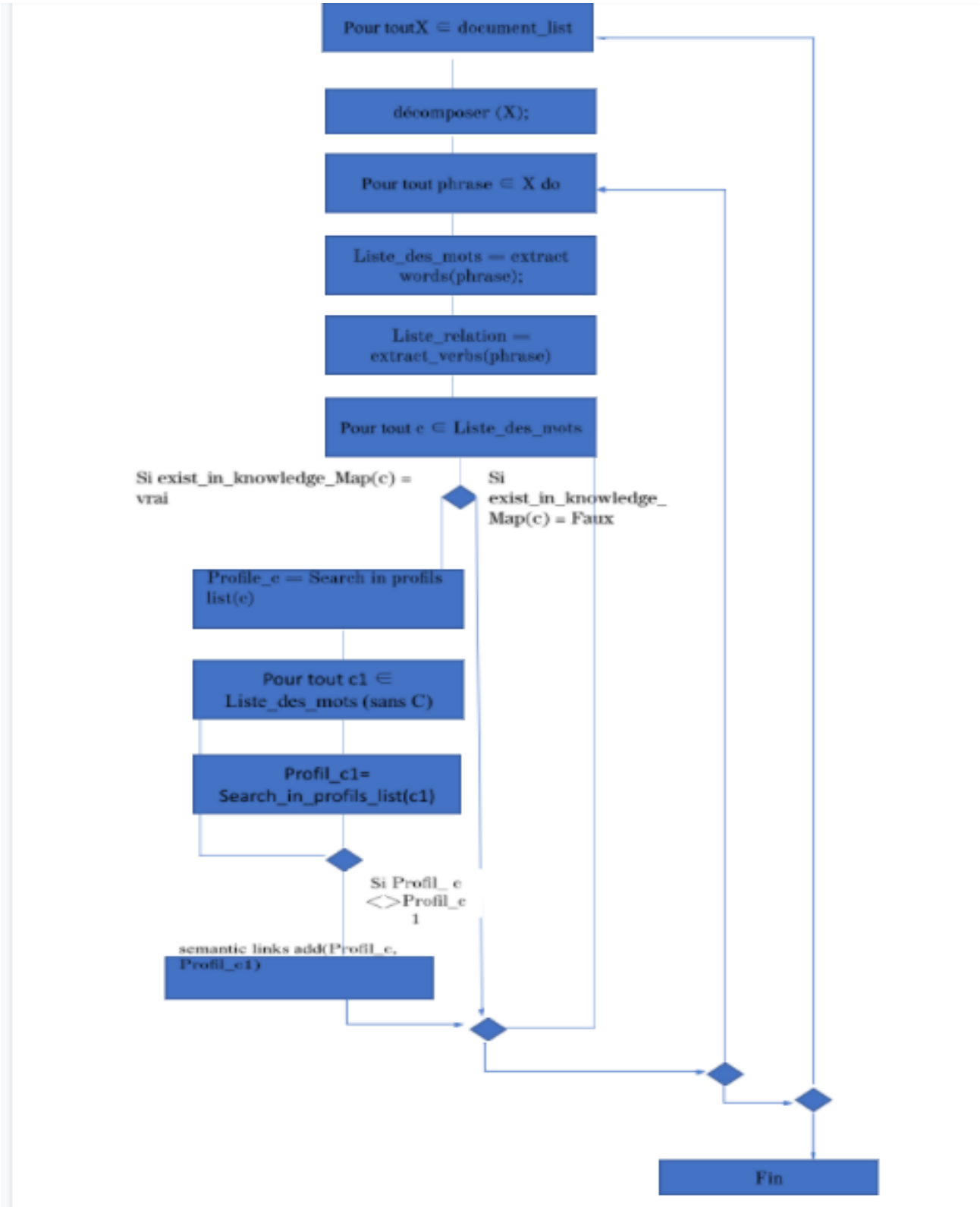


FIGURE 6.12 – Algorithme de recherche des liens sémantiques

Il existe d'autres techniques basées sur la classification supervisée pour lier sémantiquement des documents. Tsoumakas et al., dans

leur approche [Tsoumakas *et al.* 2009], ont proposé de construire un ensemble de M classificateurs où M est le nombre d'étiquettes. Chaque classificateur reconnaîtra seulement une seule étiquette. La phase d'apprentissage de chaque classificateur se fait avec un jeu de données spécifique. Le jeu de données de chaque classificateur est obtenu en transformant le jeu de données initial pour un classificateur i. Ils en construisent un nouveau composé des mêmes documents de l'ensemble des données initiales mais avec un nouvel étiquetage. Les documents étiquetés avec l'étiquette i seront étiquetés avec une étiquette P et les autres seront étiquetés avec une étiquette N avec ce nouveau jeu de données. Par la suite, le classificateur i apprendra à distinguer les documents de l'étiquette i des autres documents. Hullermeier, et al., proposent une approche différente [Hüllermeier *et al.* 2008] qui ordonne les étiquettes en fonction de leur pertinence pour un document. Les auteurs s'appuient sur une transformation appelée apprentissage par paires. Cette transformation aboutit à autant de classificateurs qu'il y a de paires d'étiquettes. Chaque classificateur fait la distinction entre deux étiquettes. L'ensemble des données ne contient que des documents étiquetés avec l'une des deux étiquettes de la paire, jamais les deux. Il construit m(m1) classificateurs où m est le nombre d'étiquettes. Hullermeier et al., [Hüllermeier *et al.* 2008] considèrent chaque prédiction d'un classificateur comme un vote pour l'une des étiquettes de la paire, donc l'étiquette avec le plus de votes est classée première, la deuxième ayant reçu le plus de votes est classée deuxième et ainsi de suite.

D'autres approches reposent à la fois sur le clustering et la classification : Smail Sellah [Sellah 2019], propose une approche en trois phases : construire une structure sémantique à partir de documents d'entreprise (map of knowledge) puis représenter les documents comme des vecteurs dans lesquels chaque élément représente une

caractéristique. Après cela, il propose d'appliquer à la fois des approches supervisées et non supervisées pour classer et regrouper les documents par thème. Construire un dictionnaire local de l'entreprise basé sur des mots extraits de documents comme proposé dans cette approche. Il est important pour identifier le lexique de l'entreprise. Cependant, cela rend la similitude entre les concepts limitée. Il est important de se référer à d'autres dictionnaires lexicaux pour trouver une similitude entre les mots.
Dans notre approche, à l'aide des profils, nous construisons des liens entre les collaborateurs de l'entreprise et les documents dont ils ont besoin.

6.4.2 La distribution des documents dans Know-linking :

6.4.2.1 Principe :

Le principe de distribution des documents dans notre approche Know-linking repose principalement sur deux principes qui sont les suivants :

— Analyser le contenu du document pour identifier le profil qui peut l'intéresser ;
— Suite à la génération des liens sémantiques, distribuer les documents contenant ces liens.

Nous modélisons le principe de distribution dans la figure 6.13 ci-dessous.

Nous procédons donc en deux étapes :

— Relier les documents aux profils :
Après la structuration des besoins des collaborateurs en supports de connaissances sous forme des profils, il est possible de chercher les documents qui les intéressent en se basant sur la représentation du profil et en comparant celle-ci à la représentation du contenu du document. En effet, chaque concept

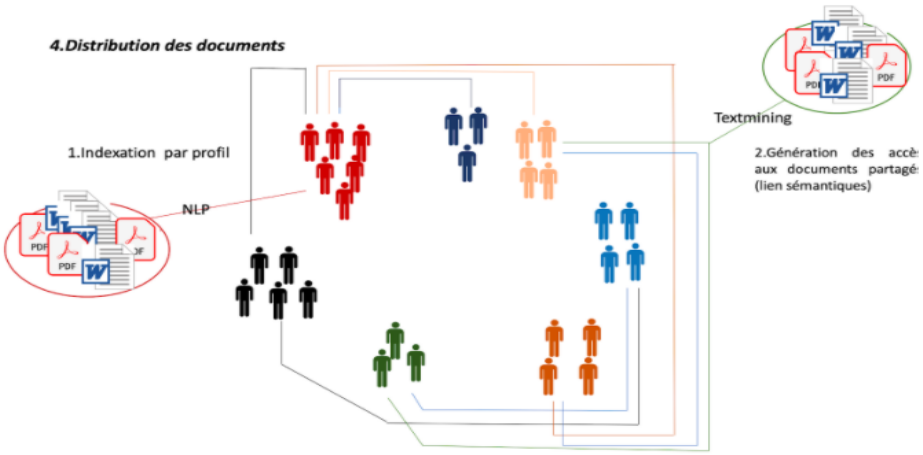


FIGURE 6.13 – Distribution des documents dans Know-linking

de la représentation des profils présente un sujet d'un document ou plus. Pour cela nous proposons dans Know-linking d'établir des liens sémantiques entre les documents et les profils. Nous modélisons le principe de cette indexation par profil dans la figure 6.14 ci-dessous.

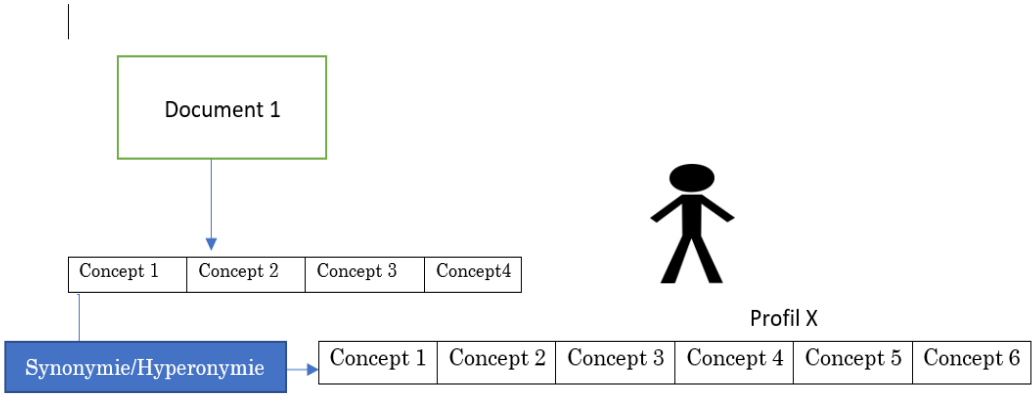


FIGURE 6.14 – Le principe de l'indexation par profil

Si le taux de correspondance entre les concepts représentant le document et le profil dépasse 50% nous indexons le document 1 sous le profil X.

— Partager l'accès aux documents contenant des liens sémantiques :

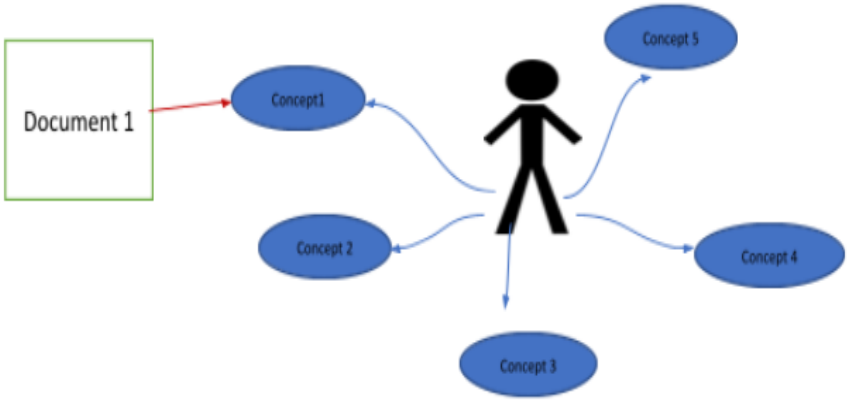


FIGURE 6.15 – Indexation par profil

Cette seconde sous-étape consiste à générer des accès particuliers aux documents contenant des liens sémantiques. Bien que ces documents doivent être partagés entre plusieurs profils, nous ne les dupliquons pas. Par contre nous permettons l'accès au document pour le profil correspondant via un lien d'accès (figure 6.16).

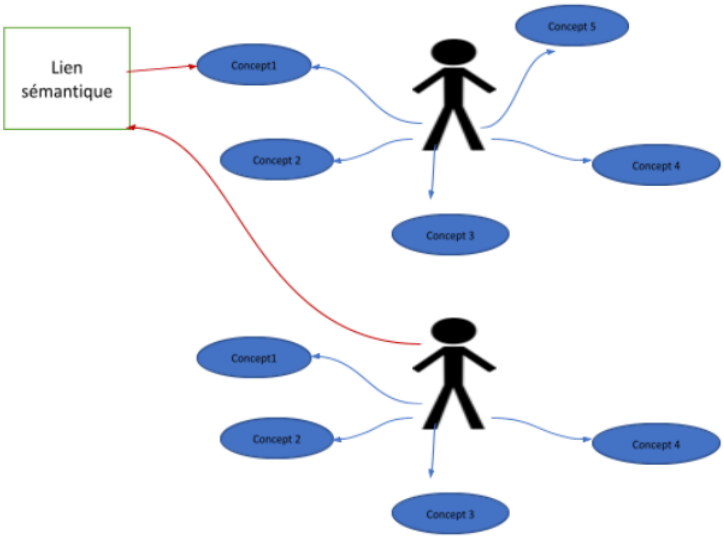


FIGURE 6.16 – Génération des accès aux documents partagés

6.5 Conclusion

Nous proposons dans notre étude de relier les supports de connaissances (les documents) avec les profils et de les rendre accessibles aux différents demandeurs. La représentation des profils sous forme de graphe permet d'une part de modéliser les concepts d'un profil ainsi que les liens entre eux. D'autre part la combinaison de l'ensemble de graphes en un seul modèle résulte en un graphe de connaissance de l'entreprise. L'approche Know-linking construit des cercles de connaissances ayant le collaborateur comme centre et où chaque profil est lié au support de connaissances sur la base de son besoin. Ce cercle s'accroît grâce à une alimentation continue par de nouveaux concepts extraits au fur et à mesure.
La nouvelle stratégie de distribution de document dans Know-linking permet au collaborateur d'être lié uniquement aux documents dont il a besoin. Cette démarche est importante dans le cycle de partage de connaissances. Cependant, ce travail de recherche peut être complété par une partie de recommandation permettant d'injecter des documents dans les outils de travail à des étapes bien précises de la réalisation des projets. Afin de valider la faisabilité des principes de l'approche Know-linking présentée dans les chapitres 5 et 6, nous avons réalisé une infrastructure logicielle respectant les trois phases. Nous détaillons la réalisation de cette infrastructure ainsi que les résultats dans la suite du rapport.

Chapitre 7

Know-linking : l'infrastructure logicielle

"Le logiciel est une excellente combinaison entre l'art et l'ingénierie." Bill Gates

Sommaire

7.1 Introduction

Les recherches que nous avons menées ont abouti à une approche ciblée de traçabilité et de page de connaissances de l'entreprise. Notre approche est composée de trois étapes :

1. **Profilage de collaborateurs :** identifier les profils des collaborateurs de l'entreprise selon leurs besoins en connaissances ;
2. **Génération des graphes et liens sémantiques entre les profils :** représenter les profils de l'entreprise sous forme d'un graphe de connaissances et étudier les liens entre eux ;
3. **Distribution des documents :** indexer les documents dans les profils correspondants et générer des accès pour les documents afin de les partager entre les collaborateurs.

L'étude théorique de l'approche Know-linking nous a permis de tracer l'architecture de l'approche en nous basant sur des spécifications fonctionnelles. On entend par l'architecture un ensemble de composants logiciels et de bibliothèques assurant les étapes définies dans l'étude théorique de l'approche. Le présent chapitre est une présentation de l'approche Know-linking d'un point de vue architectural et conceptuel. Nous commençons par identifier les spécifications puis nous montrons l'architecture générale de cette infrastructure.

7.2 Spécification et étude théorique de l'infrastructure logicielle

7.2.1 Fonctionnalités de Know-linking

La mise en place d'une solution logicielle basée sur l'approche Know-linking exige des fonctionnalités multiples pour les acteurs

dans une entreprise. Bien que Know-linking concerne toute la population de l'entreprise, ils ne bénéficient pas des mêmes fonctionnalités dans la solution. Nous commençons alors par étudier les différents rôles qu'un utilisateur peut avoir dans Know-linking :

— Administrateur : le rôle d'une personne (ou plus) qui a les droits d'administration de l'infrastructure ;

— Utilisateur : un simple utilisateur qui possède des droits basiques de consultation, d'ajout de contenu ;

— Le système : c'est le cœur de l'infrastructure, ou autrement dit l'ensemble des composants logiciels qui analysent en toute autonomie les données d'entrée. Il crée automatiquement la liste des profils, génère les graphes et les liens sémantiques, puis il distribue les documents.

7.2.2 Fonctionnalités par acteurs

Les fonctionnalités offertes par l'infrastructure qui implémente l'approche Know-linking peuvent être classées selon les rôles présentés précédemment.

7.2.2.1 L'administrateur

L'administrateur de Know-linking est le responsable du lancement de la mise à jour de l'infrastructure. Il s'agit du lancement de la mise à jour automatique de la liste des profils, des graphes des profils, du graphe de l'entreprise, des liens sémantiques et des documents distribués. Il a le droit de semi-superviser les résultats en entrant un jeu de données. Il peut également vérifier de nouveau les concepts générés ainsi que leur adéquation avec les profils. Nous nous basons sur le diagramme d'UML, notamment les cas d'utilisations, pour modéliser le rôle Administrateur avec ses différentes fonctionnalités (la figure 7.1).

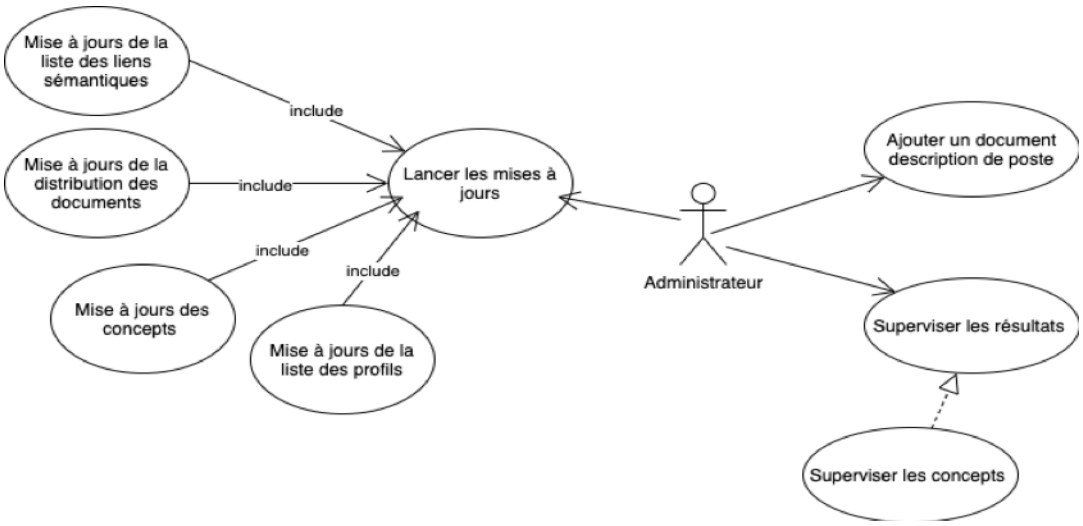


FIGURE 7.1 – Diagramme de cas d'utilisation de l'administrateur

7.2.2.2 L'utilisateur

L'utilisateur peut ajouter un ou plusieurs documents de travail. Il peut également accéder en mode lecture aux graphes des profils, aux documents par profils et au graphe de connaissances. À l'aide du diagramme de cas d'utilisation UML nous modélisons ce rôle dans la figure 7.2 ci-dessous.

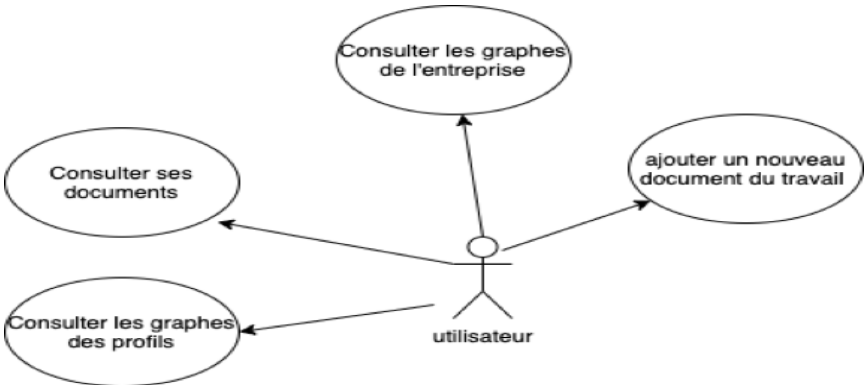


FIGURE 7.2 – Diagramme de cas d'utilisation de l'utilisateur

7.2.2.3 Acteur système

L'acteur système est le responsable de la création des profils. Il génère les graphes, les liens sémantiques et il distribue les documents selon les profils et liens entre eux. Il effectue également la

mise à jour de ces éléments : les concepts, les profils, le graphe, les liens sémantiques et la distribution des documents. Cette mise à jour s'effectue après avoir été lancée par l'administrateur. Le diagramme de cas d'utilisation UML nous permet de modéliser ses fonctionnalités dans la figure 7.3 ci-dessous.

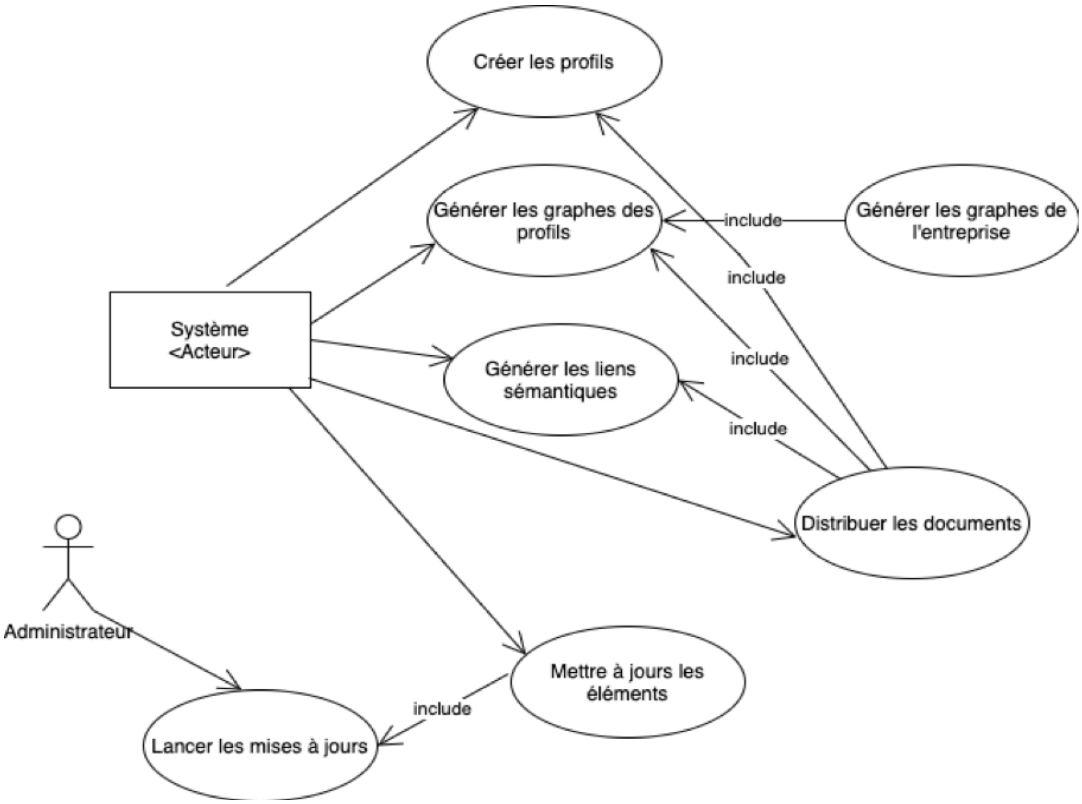


FIGURE 7.3 – Diagramme de cas d'utilisation du système

7.3 Flux de données

7.3.1 Cycle de vie des objets dans Know-linking

Les objets d'un système d'information subissent des changements durant leur cycle de vie, allant de la création à la destruction ou l'archivage. UML a mis à disposition des concepteurs le diagramme d'état-transition pour modéliser les différents états d'un objet traité par le système. Dans notre étude conceptuelle nous

considérons comme particulièrement importants deux objets que nous allons concevoir à savoir les profils et les documents.

7.3.1.1 Cycle de vie des profils

Un profil, une fois créé, va être enregistré. Les profils enregistrés subissent des mises à jour périodiquement. Ces mises à jour concernent les missions, les outils ou les concepts. Le profil, à la fin de son cycle de vie, doit être archivé pour la traçabilité.

Nous modélisons les différents états d'un document dans le diagramme d'états transitions suivant (figure 7.4).

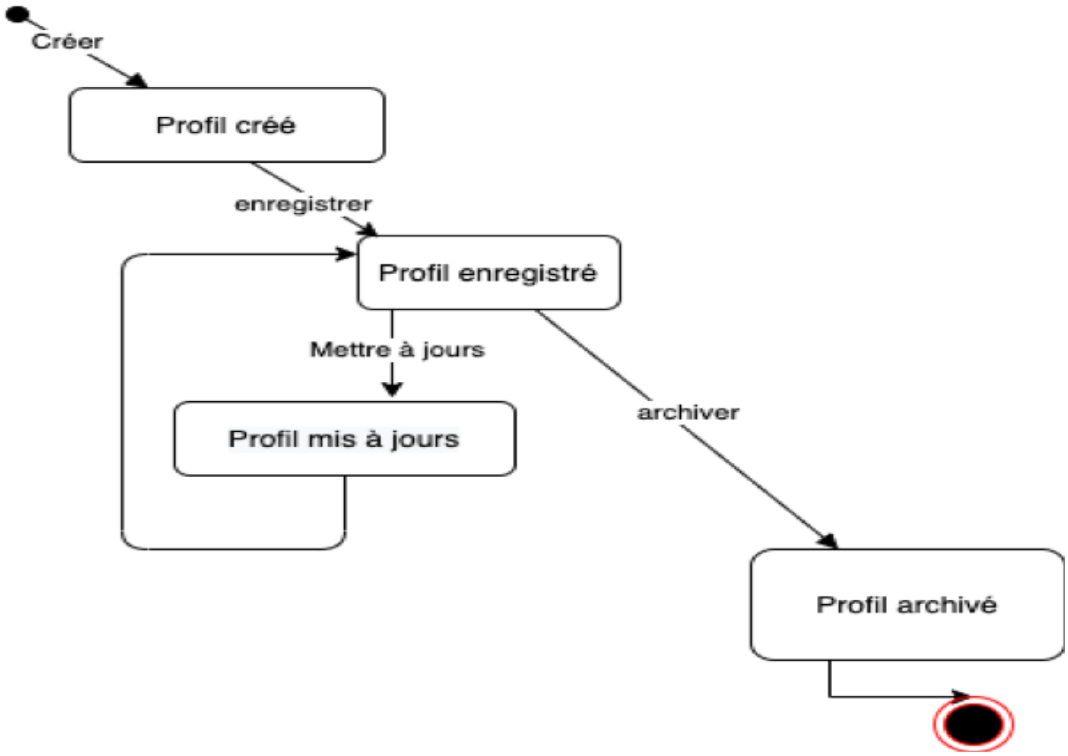


FIGURE 7.4 – Diagramme d'état transition d'un profil

7.3.1.2 Cycle de vie des documents

Un document de travail créé par un collaborateur va être analysé. Son contenu va déterminer tous les profils qui peuvent être intéressés. Au final, ce document va être indexé dans les profils identifiés.

Nous modélisons les différents états d'un document dans le diagramme d'états transitions suivant (figure 7.5).

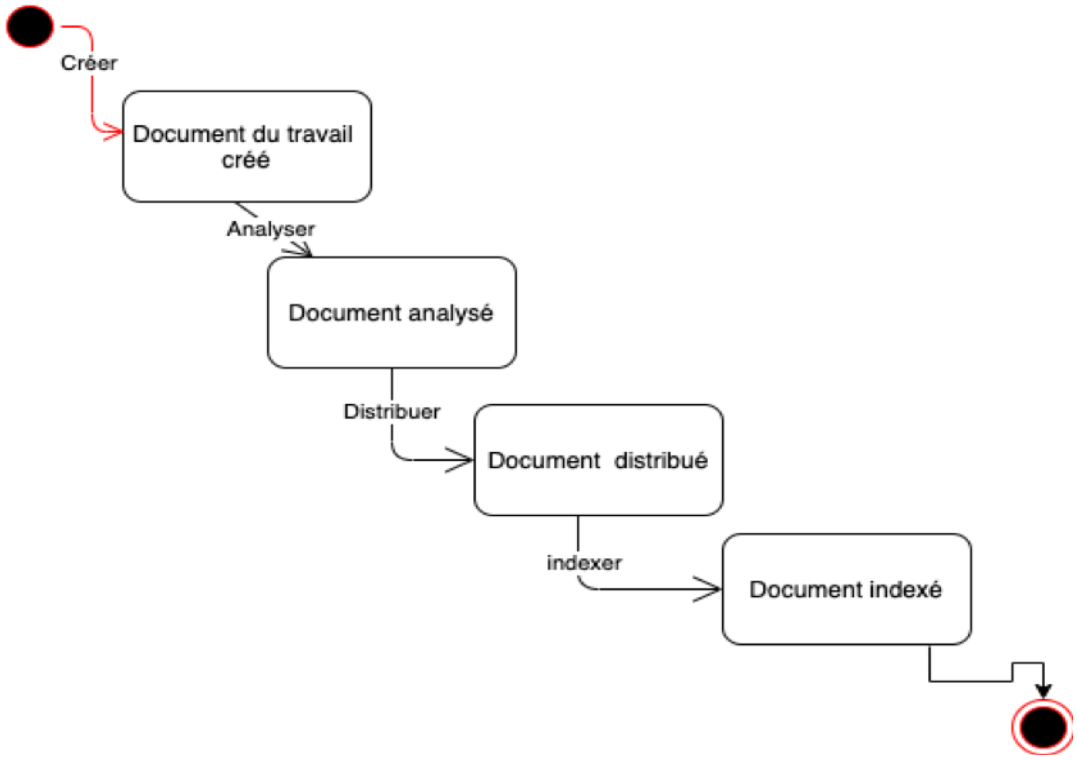


FIGURE 7.5 – Diagramme d'état transition d'un document de travail

7.4 Exécution dynamique

7.4.1 Scénario « utilisateur »

L'objectif de Know-linking est de distribuer les documents qui contiennent des connaissances partagées entre les acteurs d'une façon automatique. Cela suppose que les instructions de knwo-linking s'effectue en « backend ». Nous présentons un scénario classique qui commence par la création d'un document de travail d'un collaborateur en supposant que les profils et que les graphes ont déjà été créé, et qui se termine par la redistribution de ce document entre les collaborateurs concernés.

1. Un utilisateur crée un document dans son interface ;

2. L'analyseur du contenu charge le document ;
3. L'analyseur du contenu applique la méthode Bag of concept sur ce document ;
4. Les résultats du bag of concepts seront envoyés au distributeur des documents ;
5. Le distributeur des documents cherche la similarité des concepts des documents par rapport aux concepts des profils ;
6. Le distributeur identifie le profil correspondant ;
7. Si le composant distributeur des documents détecte un nouveau concept il l'ajoute à la liste du profil identifié et le met à jour ;
8. Suite à la mise à jour, un nouveau graphe de connaissances est généré par le composant générateur des graphes ;
9. Le générateur des liens sémantiques cherche les liens sémantiques dans ce document ;
10. Le document sera indexé au profil correspondant et sera partagé entre plusieurs profils selon les liens sémantiques trouvés ;
11. Les utilisateurs concernés seront notifiés qu'un nouveau document a été ajouté dans leurs listes de documents.

Nous utilisons pour la modélisation de ce scénario le diagramme de séquence du langage UML (figure 7.6).

7.4.2 Scénario « administrateur »

Un administrateur est le responsable de lancement de Know-linking pour mettre à jour la distribution des documents après la génération des profils (pour prendre en considération les nouveaux profils), la génération des graphes et des liens sémantiques. Considérons alors le scénario suivant :

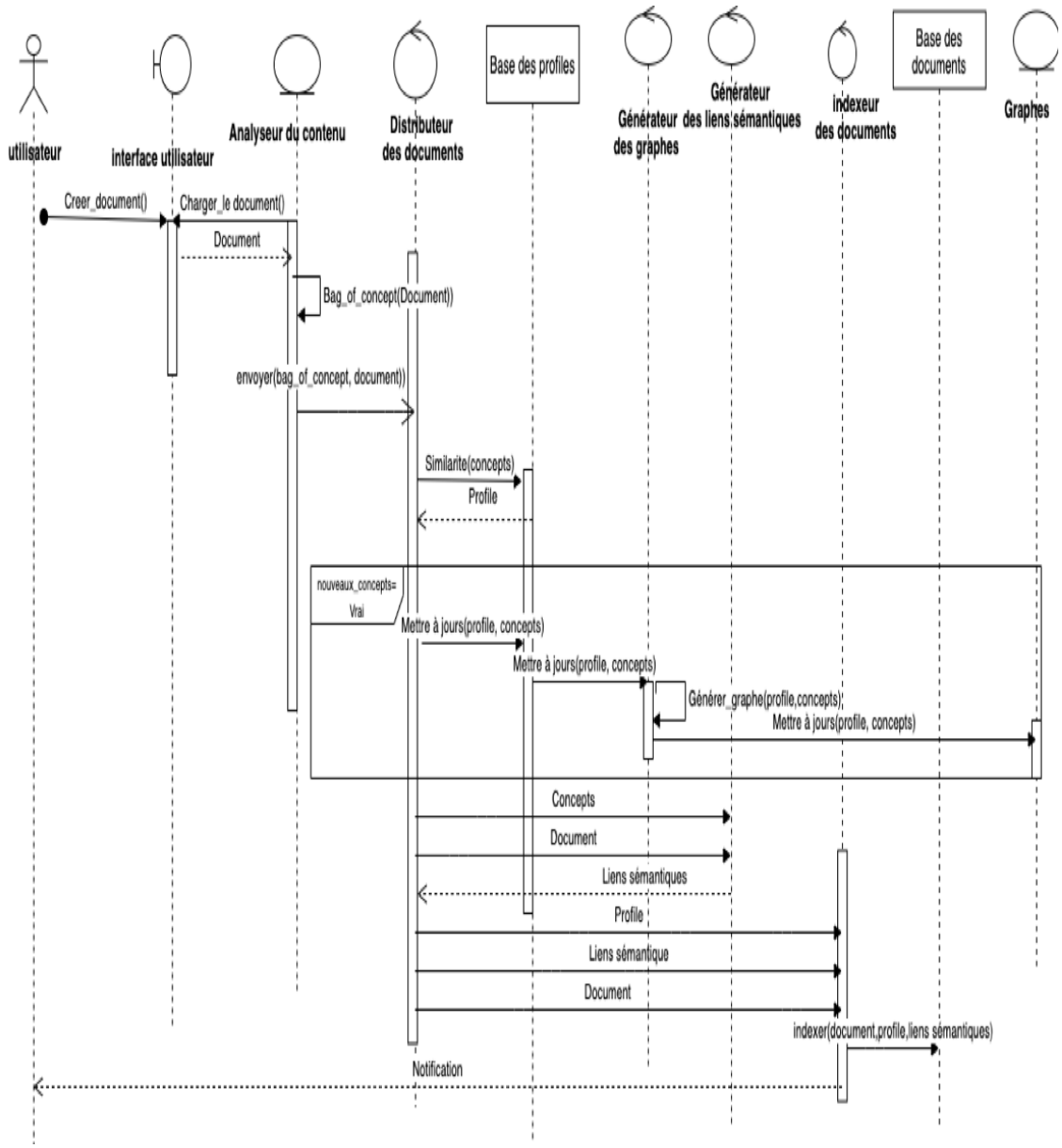


FIGURE 7.6 – Diagramme de séquence "distribuer un document"

1. L'administrateur lance une mise à jour périodique à travers son interface ;
2. Le générateur des profils reçoit un signal de lancement des mises à jour ;
3. Le générateur des profils procède à l'analyse des documents ressources humaines ;
4. Le générateur des profils procède à l'extraction des données de gestion de projets ;

5. La liste des profils extraite est envoyée par le générateur des profils à la base liste des profils pour sauvegarde ;
6. La nouvelle liste est sauvegardée ;
7. La méthode bag of concepts est appliquée sur les éléments des profils ;
8. Le générateur des graphes génère des graphes de profils ;
9. La distribution des documents est lancée une fois les graphes obtenus ;
10. Le distributeur des documents demande au générateur des liens sémantiques qui permettent d'identifier les liens sémantiques sur les ressources des profils ;
11. Les documents sont analysés et les liens sémantiques sont détectés ;
12. La distribution des documents commence par le distributeur sur la base des profils et des liens identifiés ;
13. Les documents sont indexés aux différents profils ;
14. Tous les utilisateurs sont notifiés de la nouvelle mise à jour.

Ce processus de mise à jour est modélisé à l'aide du diagramme de séquence du langage UML dans la figure 7.7 ci-dessous.

7.5 Les exigences de la mise en place de l'infrastructure Know-linking

Nous avons défini un ensemble d'exigences fonctionnelles et d'autres techniques que la mise en place d'une infrastructure logicielle basée sur l'approche Know-linking doit respecter.
Nous présentons l'ensemble des exigences dans les tableaux 7.1,7.2 et 7.3 suivants organisés par étape de l'approche Know-linking.

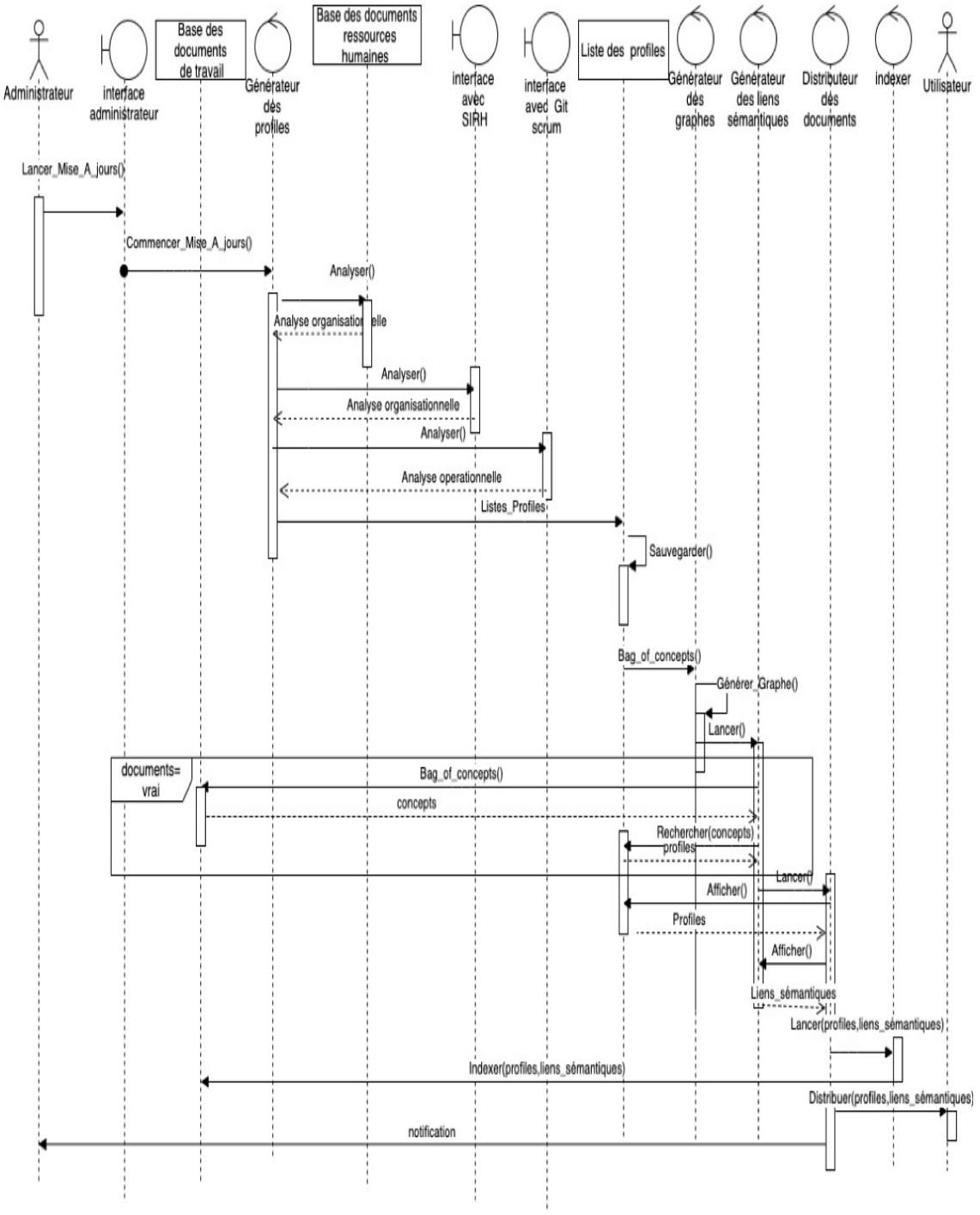


FIGURE 7.7 – Diagramme de séquence " mise à jour "

7.5.1 Génération des profils

Les exigences que nous avons spécifiées pour la génération des profils sont présentées dans le tableau 7.1.

TABLE 7.1 – Les exigences de la génération des profils

Code	Exigence	Type d'exigence
01	L'infrastructure doit analyser des données extraites à partir des outils de gestion de projets ;	Exigences fonctionnelles
02	L'infrastructure doit analyser des données extraites à partir des outils de gestion des ressources humaines ;	
03	L'infrastructure doit mettre à jour dynamiquement la liste des profils ;	
04	L'infrastructure ne doit pas générer deux profils semblables ;	
05	L'infrastructure doit combiner les résultats de deux analyses, opérationnelle et organisationnelle, pour la génération des profils ;	
06	La génération des profils de collaborateurs doit être automatique ;	
07	La génération des profils de collaborateurs par Know-linking doit être semi-supervisé ;	
08	Les profils générés par Know-linking doivent être sauvegardés ;	
09	L'analyse des outils de gestion des ressources humaines dans Know-linking se base sur l'interfaçage par des requêtes sql avec la base de données du système en question ;	Exigences Techniques
10	L'analyse des outils de gestion des projets dans Know-linking se base sur l'interfaçage par des requêtes de base de données du système en question ;	
11	L'infrastructure doit suivre l'algorithme de génération des profils (défini dans le chapitre 6) ;	
12	L'infrastructure Know-linking doit analyser les descriptions des postes en se basant sur les règles définies dans le chapitre 5 ;	
13	L'infrastructure Know-linking doit analyser les descriptions des postes à l'aide d'une bibliothèque de ressources linguistiques comme WordNet, pour détecter la synonymie et l'hyperonymie ;	
14	L'infrastructure Know-linking analyse les descriptions des postes à l'aide d'une bibliothèque de Natural langage processing permettant le raisonnement sémantique ;	

7.5.2 Génération des graphes des profils et des liens sémantiques

Les exigences que nous avons spécifiées pour la génération des graphes des profils et des liens sémantiques sont présentées dans le tableau 7.2

TABLE 7.2 – Les exigences de la génération des graphes et des liens sémantiques

Code	Exigence	Type d'exigence
01	L'infrastructure logicielle doit générer automatiquement des représentations de profils sous forme de graphes de connaissances ;	Exigence fonctionnelle
02	L'infrastructure logicielle ne doit considérer que les concepts lors de la modélisation des graphes ;	Exigence fonctionnelle
03	La mise à jour des graphes se fait d'une façon automatique une fois les mises à jour lancées par l'entreprise ;	Exigence fonctionnelle
04	Les concepts liés par un verbe dans une phrase doivent être liés dans le graphe ;	Exigence technique
05	Les liens sémantiques doivent être générés automatiquement ;	Exigence fonctionnelle
06	Les liens sémantiques doivent considérer les relations entre les termes dans le contenu textuel	Exigence technique
07	La mise à jour des liens sémantiques doit être lancée automatiquement lors du lancement des mises à jour par l'entreprise ;	Exigence fonctionnelle
08	L'infrastructure pour la découverte des liens sémantiques doit implémenter l'algorithme défini dans le chapitre 7 ;	Exigence technique
09	L'infrastructure doit considérer les relations de synonymie et hyperonymie lors de l'application de l'algorithme de génération des liens sémantiques ;	Exigence technique

7.5.3 Distributions des documents

Les exigences que nous avons spécifiées pour la distribution des documents sont présentées dans le tableau 7.3

7.6 Étude technique de l'infrastructure

L'architecture logicielle que nous présentons dans cette partie est basée sur une modélisation « basée sur des composants » dans laquelle l'infrastructure logicielle de Know-linking est vue comme un ensemble de composants inter-opérable où chacun doit réaliser le traitement d'une partie précise. Nous représentons cette architecture dans la figure 7.8 ci-dessous.

TABLE 7.3 – Les exigences de la distribution des documents

Code	Exigence	Type d'exigence
01	L'infrastructure logicielle qui implémente les principes de Know-linking doit distribuer dynamiquement les documents aux différents profils	Exigence fonctionnelle
02	La distribution des documents doit respecter les principes définis dans le chapitre 7.	Exigence fonctionnelle
03	La mise à jour des graphes se fait d'une façon automatique une fois les mises à jour lancées par l'entreprise ;	Exigence fonctionnelle
03	Les profils doivent avoir accès aux documents partagés ;	Exigence fonctionnelle
04	Les documents partagés ne doivent pas être dupliqués ;	Exigence technique

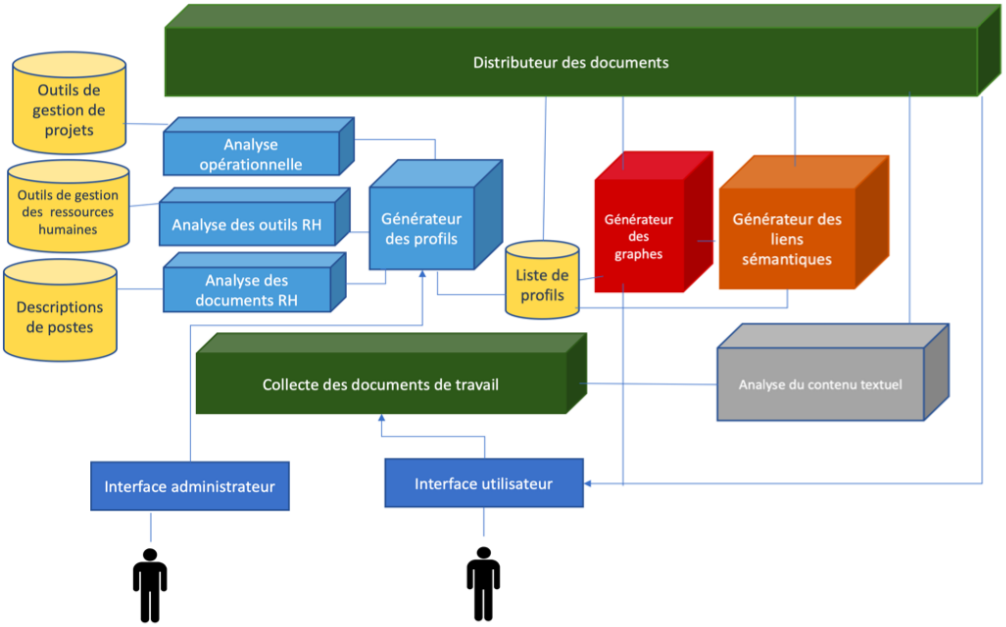


FIGURE 7.8 – Architecture Know-linking

7.6.1 Les composants logiciels de Know-linking

1. **Front-end**

   L'interface utilisateur

   L'interface utilisateur est une fenêtre qui sert à collecter le contenu créé par les utilisateurs en temps réel. Cette interface doit être intégrée à d'autres systèmes de gestion documentaire ou de gestion de connaissances où les documents

s'ajoutent dynamiquement pour être indexés ou archivés. À travers cette interface, un utilisateur peut ajouter directement un nouveau document pour qu'il soit distribué. Une option qui s'ajoute à cette interface est la visualisation des graphes des profils.

L'interface administrateur À travers cette interface l'administrateur peut lancer les mises à jour de la distribution des documents. Cette interface d'administration lui offre aussi la possibilité de considérer des jeux de données pour superviser la génération des profils de l'entreprise.

2. **Back-end**

   L'entrée de l'infrastructure logicielle

   L'entrée de l'infrastructure logicielle des outils de gestion des projets est un ensemble de données provenant de l'interfaçage avec les outils de gestion de projets, les outils des ressources humaines et les descriptions des postes. Ces données seront analysées (comme expliqué dans le chapitre précédent) par les trois composants d'analyse opérationnelle, des outils ressources humaines et des documents ressources humaines (précisément les descriptions de poste). Cette partie est identifiée dans la figure 7.9 ci-dessous.

   — Le composant analyse opérationnelle : c'est un composant assurant l'interfaçage avec les bases de données des outils de gestion de projets tel que GitScrum, Jira et autres. Afin d'extraire des informations essentielles pour la construction des profils telles que les tâches et les projets dans lesquels le collaborateur est impliqué selon le principe que nous avons décrit dans le chapitre 6 (génération des profils).

   — Le composant analyse des outils ressources humaines : c'est un composant logiciel permettant l'interfaçage avec

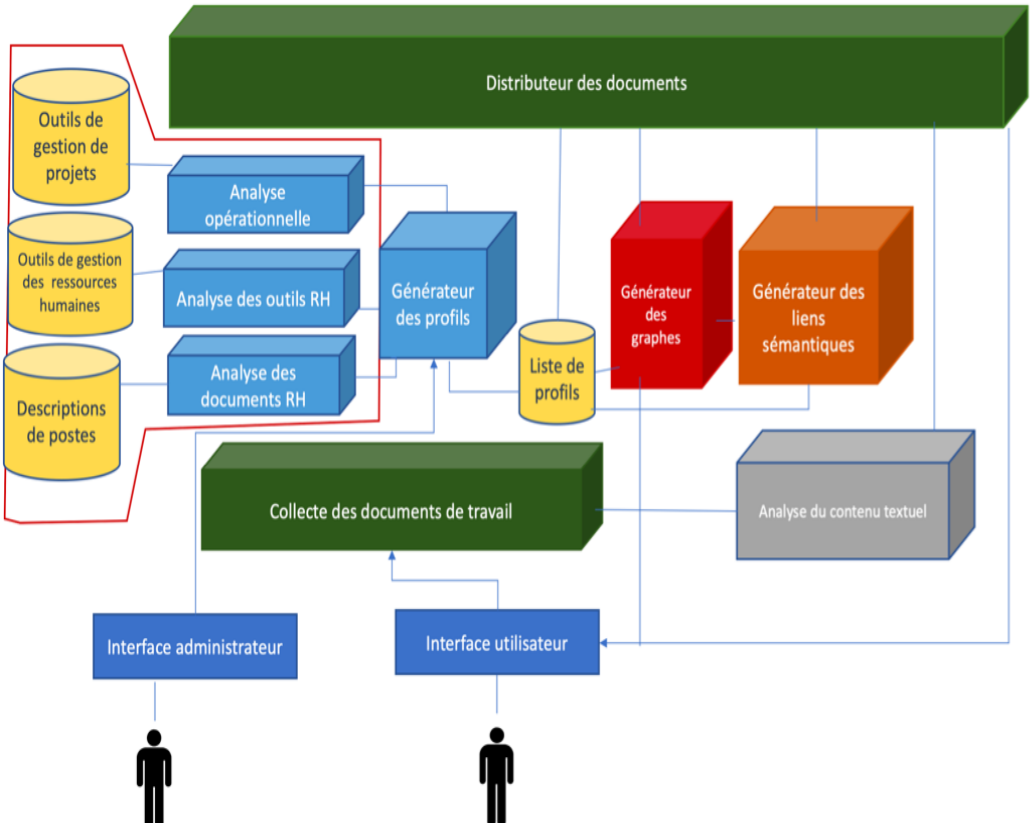


FIGURE 7.9 – La partie entrée de l'architecture

les bases de données des outils de ressources humaines afin d'extraire les données nécessaires à l'identification de la position du collaborateur dans l'organisation (telles que son affiliation, son département et ses missions)

— Le composant analyse des documents ressources humaines : Un composant logiciel assurant l'analyse des descriptions des postes en se basant sur l'ensemble de règles (patterns) présentés dans le chapitre 5, et en utilisant le Natural language processing.

Cœur de l'infrastructure

Le cœur de l'infrastructure est l'ensemble des composants qui assurent plusieurs étapes de l'approche Know-linking (figure 7.10).

Nous détaillons les composants de la manière suivante :

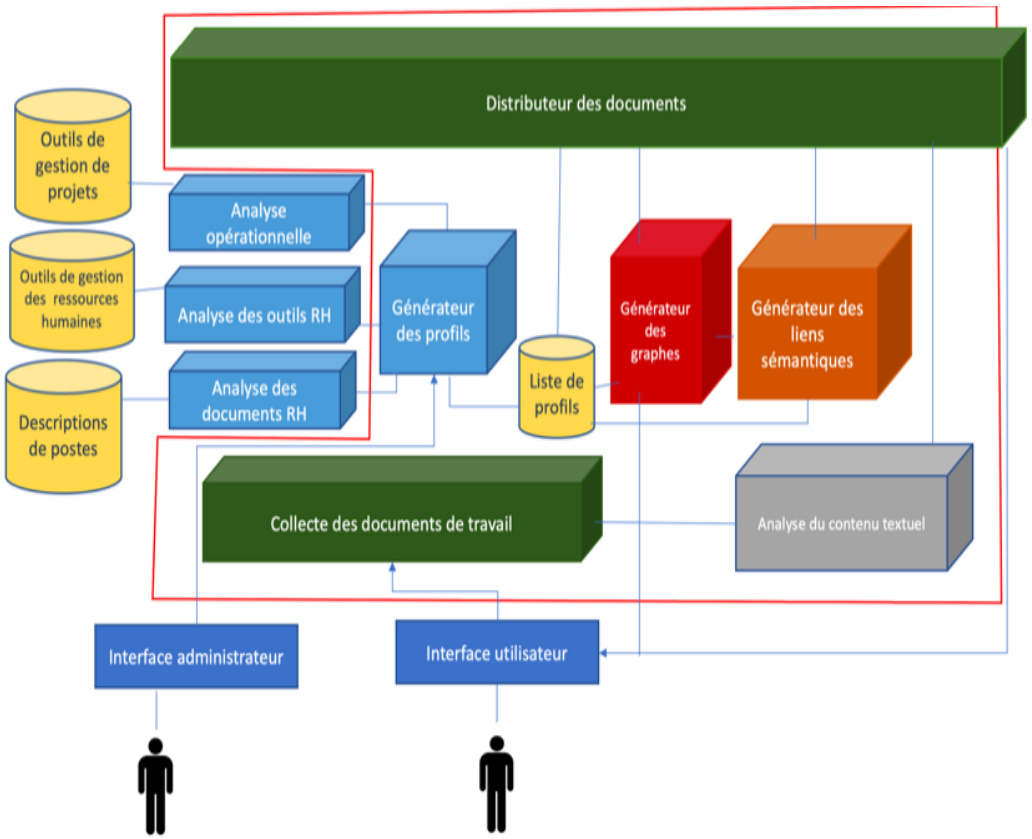


FIGURE 7.10 – Le coeur du framework

— Le générateur des profils de collaborateurs en se basant sur les données d'entrées : ce composant doit respecter les principes et les dimensions décrits dans le chapitre 6 génération des profils, tels que l'automatisation et la semi-supervision. Ce composant s'exécute selon l'algorithme défini dans le chapitre 6, en interaction avec des bibliothèques linguistiques comme WordNet.

— Liste de profils : un support de stockage de la liste des profils générés.

— Générateur des graphes : ce composant a pour vocation de transformer les profils sauvegardés dans la liste en des représentations de graphes de connaissances en se basant sur les principes définis dans le chapitre 6.

— Générateur des liens sémantiques : Ce composant procède à l'analyse des documents pour trouver les liens sémantiques

cachés entre les profils. Il implémente l'algorithme de recherche de liens sémantiques que nous avons défini dans le chapitre

— Analyse de contenu : L'analyse de contenu englobe le nettoyage des documents et l'implémentation de la méthode bag of concepts, les bibliothèques du traitement automatique des langues naturelles et les ressources linguistiques utilisées pour l'analyse, comme la bibliothèque WordNet.

— Collecte des documents de travail : ce composant a pour mission de détecter si un nouveau document a été créé afin de lancer l'analyse de son contenu.

— Distributeur des documents : ce composant prend en charge la distribution des documents selon les profils et les liens sémantiques entre eux. Il indexe les documents relatifs à chaque profil et génère des accès pour les documents partagés.

7.7 Conclusion

Une infrastructure logicielle (framework), qui implémente l'approche Know-linking dans ses trois étapes, doit suivre certaines exigences techniques et fonctionnelles pour la réussite du partage des connaissances.

Nous avons consacré le présent chapitre à l'élaboration du plan d'une infrastructure logicielle qui implémente l'approche que nous avons proposée, en suivant les spécifications et les exigences définies afin d'assurer un partage dynamique des supports de connaissances de l'entreprise.

L'étude théorique de l'infrastructure logicielle « Know-linking » est suivie d'un choix technique pour sa réalisation, d'un travail de développement pour tester sa faisabilité ainsi que des résultats de

l'approche sur un terrain réel. Nous détaillons cette partie dans le chapitre suivant.

Chapitre 8

Know-linking sur le terrain : Expérimentation

C̈'est souvent gratifiant d'expérimenter et constater les fruits d'avoir osé. ¨Geneviève Krebs

Sommaire

8.1 Introduction

L'approche Know-linking est le fruit d'un travail de recherche scientifique et d'études bibliographiques en réponse à des problématiques de recherche que nous avons posées au début de ce rapport. Ces efforts ont permis d'établir une approche basée sur des principes de profilage et de partage continu de connaissances. Afin de prouver la faisabilité de notre approche et la qualité des résultats, nous avons pensé à mettre en place une infrastructure logicielle (framework) appliquant l'approche Know-linking, dont nous avons rédigé les spécifications et les exigences dans le chapitre 7. Nous avons également réalisé un plan de tests de l'infrastructure que nous avons implémentée pour comparer les résultats obtenus par rapport à ceux attendus. Nous détaillons dans le présent chapitre le terrain d'application choisi, le protocole d'expérimentation qui englobe les choix techniques que nous avons effectués ainsi que les démarches de développement. Nous présentons en outre les résultats que nous avons obtenus.

8.2 Terrains d'application et audit

Nos terrains d'application se situent dans deux domaines différents : l'industrie de l'aéronautique et le secteur de l'énergie. Nous présentons dans ce chapitre les résultats de l'aéronautique en particulier.

Pour pouvoir comprendre le terrain et avoir plus de visibilité sur le besoin réel de l'application de Know-linking, nous avons réalisé un travail d'audit technique par le biais d'entretiens avec les collaborateurs de l'entreprise. Cette étape constituait une base pour la compréhension des flux de données, des processus métiers, des supports de connaissances existants, des méthodes de traçabilité (si

elles existent), mais aussi des systèmes d'information et de leurs interconnexions. Théoriquement les deux terrains sont différents mais concrètement tous deux sont dotés des mêmes caractéristiques :

1. Un nombre important de collaborateurs : dans la grande industrie, le nombre d'employés est de plusieurs dizaines de milliers ;
2. Des processus métiers complexes : les projets sont de longues durées et englobent des intervenants de plusieurs spécialités comme pour l'aéronautique : l'électrique, la mécanique, la peinture, l'informatique, l'aérodynamique, etc ;
3. Une production importante de documents : le nombre de documents produits est très important dans les différentes spécialités au quotidien ;
4. Un manque de traçabilité : nous avons remarqué après l'audit technique qu'il n'existe aucune méthode de traçabilité pertinente des connaissances ;
5. Une absence d'approche ou de méthode de gestion de connaissances partagées : les collaborateurs sont parfois obligés de refaire des études déjà existantes et réalisées par leurs collègues par manque d'information de l'existence de ces dernières.
6. Du bruit : les collaborateurs ont accès à des documents qui ne les intéressent pas ;
7. Un environnement technique classique : les solutions logicielles adoptées sont classiques et n'offrent pas d'options sophistiquées dans la gestion des documents tels que des options de recommandations ou de recherche semi-automatique. Après avoir analysé les résultats de l'audit technique, nous

avons identifié le besoin réel d'adopter une approche systémique assurant un partage de connaissances ciblées entre les acteurs de l'entreprise.

8.3 Protocole d'expérimentation

8.3.1 Choix techniques

Il existe plusieurs technologies sur le marché qui peuvent répondre aux spécifications et aux exigences de l'infrastructure Know-linking détaillées dans le chapitre 7. Pour comprendre d'une façon plus claire les composants logiciels dont nous aurons besoin pour notre approche, nous présentons dans le schéma suivant une illustration (Figure 8.1).

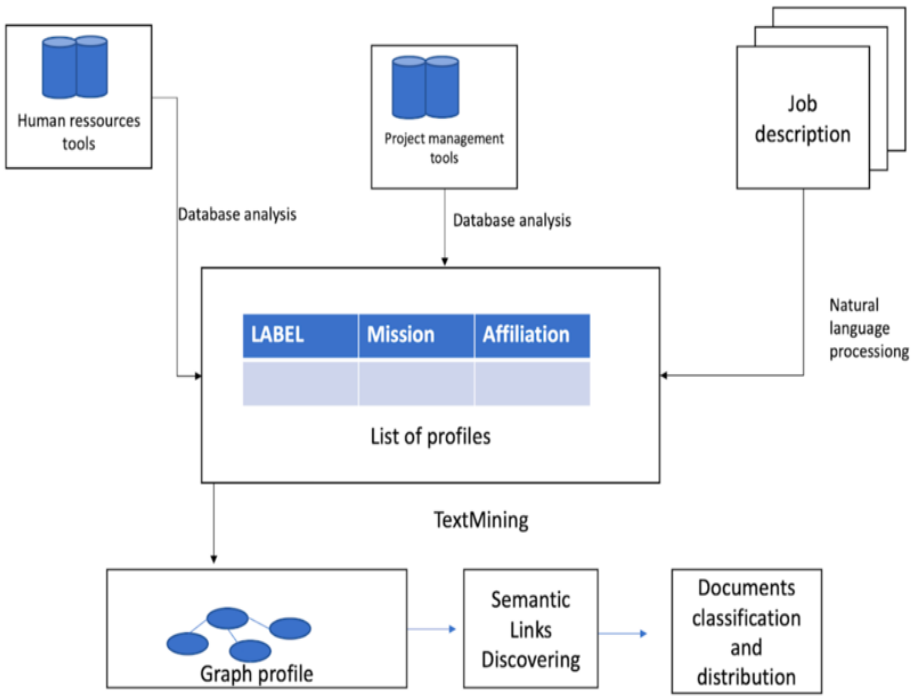


FIGURE 8.1 – Architecture know-linking

En se basant sur le modèle que nous illustrons dans la figure 8.1 nous avons choisi certains outils, langages et plateformes par

facilité d'implémentation pour l'environnement technique du démonstrateur (tableau 8.1).

TABLE 8.1 – L'environnement technique.

Base de données	MySQL
Environnement de développement intégré	Spyder, Anaconda
Langage	Python
TextMining, apprentissage automatique	Porte, textract
Graphique	Matplotlib
Traitement du langage naturel	Nltk

8.4 Interfaces

L'interface utilisateur permet à un utilisateur de consulter les graphes de connaissances et la liste de ses documents. Nous illustrons les interfaces dans le schéma 8.2 suivant.

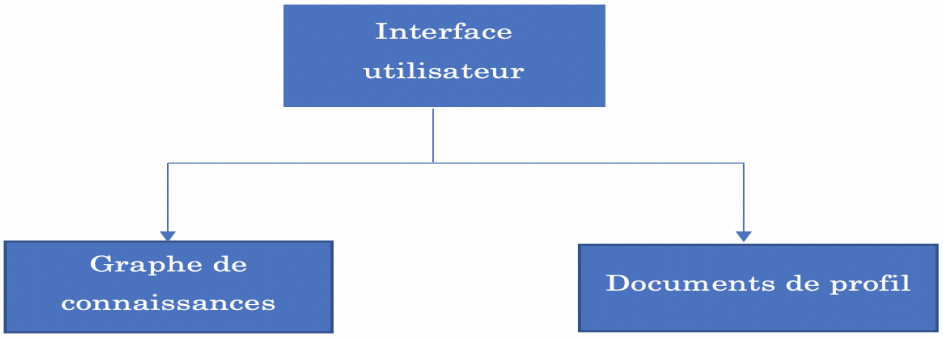


FIGURE 8.2 – Interfaces utilisateur

Quand un administrateur accède à son interface pour lancer les mises à jours des profils, le système prend par la suite le relais et met à jour les concepts, les graphes et la distribution des documents. Il a aussi la possibilité de vérifier des résultats ou de mettre un nouveau document pour être distribué. Nous présentons les interfaces d'un administrateur que nous avons développées dans le schéma ci-dessous 8.3.

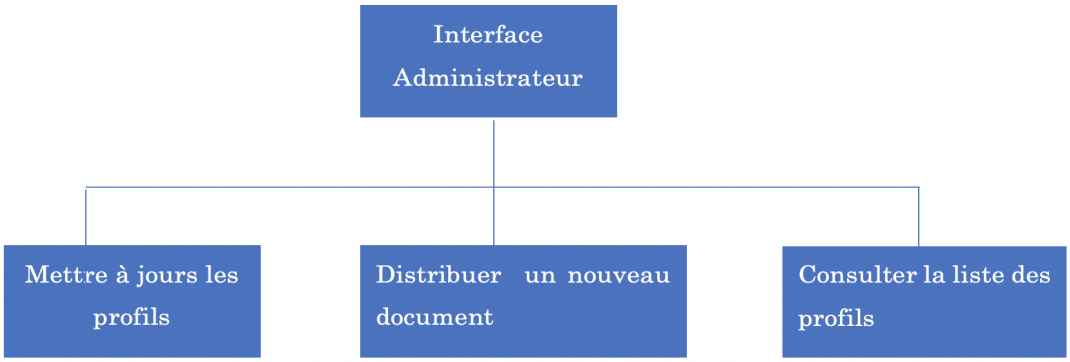


FIGURE 8.3 – Interface administrateur

8.5 Objectifs et plan de tests

Les tests que nous avons lancés à travers cette infrastructure logicielle (Framework) et qui implémentent l'approche Know-linking, suivent un plan de tests organisé. Nous détaillons tout d'abord les objectifs des tests fonctionnels dans le tableau 8.2.

TABLE 8.2 – Les objectifs des tests.

Phase du test de Know-linking	Objectifs
Les tests de l'étape génération des profils	Tester la capacité du framework à générer automatiquement des profils ;
	Vérifier la non- redondance des profils
	Tester la capacité du framework d'analyser des documents de description de poste (job description)
Les tests de l'étape génération des graphes	Tester la fonctionnalité génération des graphes des profils
	Vérifier la non-redondance des concepts dans le même graphe
	Vérifier que les graphes correspondent bien aux intitulés des profils
Les tests de l'étape génération des graphes	Tester la capacité de l'infrastructure à générer automatiquement des concepts pour chaque profil
	Vérifier qu'il n'y a pas deux concepts hyperonymes ou synonymes générés
	Vérifier que les concepts décrivent bien les profils correspondants
Les tests de la génération des liens	Tester la fonctionnalité « génération des liens sémantiques »

Phase du test de Know-linking	Objectifs
	Vérifier que les liens générés représentent bien des liens entre deux profils différents
	Tester l'efficacité de la méthode Bag-of concepts sur les documents de profils
Tests de la distribution des documents aux profils	Tester la capacité de l'infrastructure de distribuer automatiquement des documents aux profils
	Tester la capacité d'indexer des documents par profils
	Vérifier que la distribution est faite en se basant sur les profils et les liens entre eux

Les dimensions que nous avons définies pour l'approche know-linking doivent aussi être testées en tant que tests non-fonctionnels. Nous entendons par ces dimensions les réflexions que nous avons présentées lors de la partie état de l'art telles que : l'automatisation, le dynamisme et la semi-supervision.
Concernant les plans de tests, nous nous basons sur deux scénarios décrits dans les spécifications : nous les appelons respectivement le scénario utilisateur et le scénario administrateur.

Le scénario utilisateur comporte les étapes suivantes :

1. Un utilisateur ajoute un document pour distribution ;
2. L'utilisateur consulte la liste des documents du profil ;
3. L'utilisateur voit qu'un nouveau document est ajouté à son profil.

Le scénario administrateur comporte l'ensemble des étapes suivantes :

1. L'administrateur lance la création d'un nouveau profil à partir d'un document structuré ;
2. L'administrateur consulte la liste des profils ;
3. L'administrateur modifie la liste de concept d'un profil ;

4. L'administrateur consulte les résultats de la distribution des documents.

8.6 Données de test

Les données de test que nous introduisons au Framework Know-linking consistent en un ensemble de 1000 documents dans l'aéronautique. Ces documents se composent de documents ressources humaines tels que des descriptions de postes, d'autres appartiennent à des domaines en lien avec l'activité aéronautique tels que l'informatique, l'électricité, la mécanique, l'aérodynamique, etc.
Ces documents ont été entrés au fur et à mesure. Les résultats sont évalués en parallèle. L'objectif de la collecte de ces données est d'appliquer l'approche know-linking dans un contexte d'entreprise de grande envergure, multidisciplinaire et de projets de haute complexité.

8.7 Implémentation des tests

À ce niveau, l'infrastructure a été développée avec tous ces composants en respectant les spécifications définies dans le chapitre précédent. Nous commençons les scénarios de tests utilisateur et administrateur comme détaillé dans l'étape suivante afin d'atteindre les objectifs de la phase de test. Nous commençons par le premier scénario administrateur et pour chaque sous-étape de test nous montrons l'interface et l'affichage concerné.
Le scénario administrateur comporte l'ensemble des étapes suivantes :

1. L'administrateur lance la création d'un nouveau profil à partir d'un document structuré.

En principe, la génération des profils est effectuée automatiquement par le système, cependant, nous donnons la possibilité à un administrateur de générer lui aussi des profils à travers une interface lui permettant d'entrer manuellement l'ensemble des documents (opérationnel et organisationnel). Suite à cette étape, le système prend le relais et génère le profil. Par la suite il enchaine de manière autonome sur les autres étapes de know-linking : la génération des graphes, des liens sémantiques et finalement la redistribution des documents en considérant les nouveaux profils créés.
Nous présentons l'interface que nous avons développée et qui donne la possibilité à l'administrateur d'effectuer cette étape d'ajout d'un nouveau profil dans la figure 8.4.

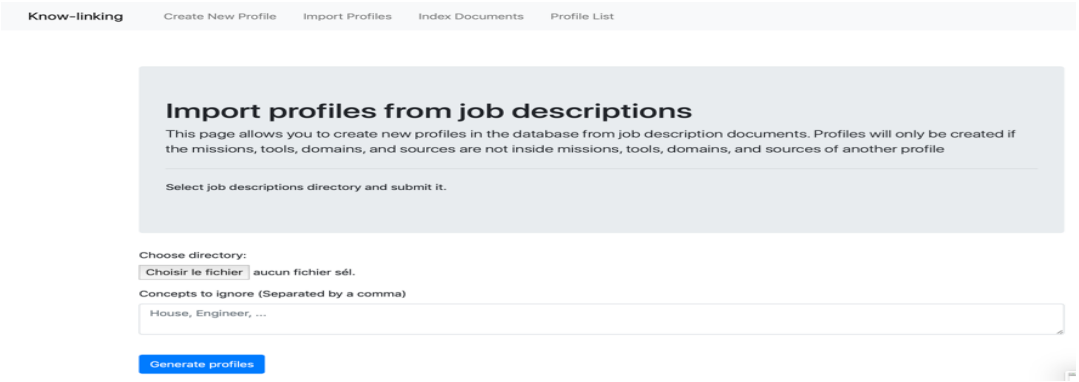
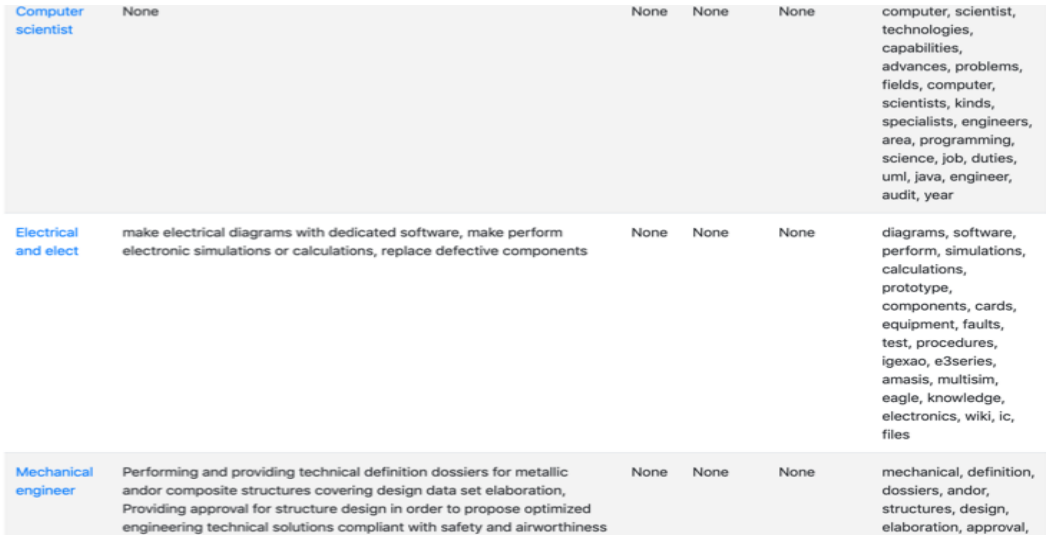


FIGURE 8.4 – Générer un nouveau profil à partir d'une description de poste

Nous précisons que le ou les documents entrés par l'administrateur seront analysés avec le Natural Langage processing. Après le nettoyage des documents nous nous basons sur des modèles linguistiques (patterns) pour l'extraction de l'information que nous avons définis dans le chapitre trois pour analyser les documents.

2. L'administrateur consulte la liste des profils
   Après avoir généré un nouveau profil, le framework va lancer

une mise à jour de l'ensemble des composants pour regénérer des résultats (graphes, liens sémantiques et distribution). L'administrateur doit alors consulter la liste des profils pour vérifier si le nouveau profil a été ajouté ou non. Nous présentons dans la figure 8.5 ci-dessous.



Computer scientist	None	None	None	None	computer, scientist, technologies, capabilities, advances, problems, fields, computer, scientists, kinds, specialists, engineers, area, programming, science, job, duties, uml, java, engineer, audit, year
Electrical and elect	make electrical diagrams with dedicated software, make perform electronic simulations or calculations, replace defective components	None	None	None	diagrams, software, perform, simulations, calculations, prototype, components, cards, equipment, faults, test, procedures, igexao, e3series, amasis, multisim, eagle, knowledge, electronics, wiki, ic, files
Mechanical engineer	Performing and providing technical definition dossiers for metallic andor composite structures covering design data set elaboration, Providing approval for structure design in order to propose optimized engineering technical solutions compliant with safety and airworthiness	None	None	None	mechanical, definition, dossiers, andor, structures, design, elaboration, approval,

FIGURE 8.5 – Générer un nouveau profil à partir d'une description de poste

3. L'administrateur modifie la liste de concept d'un profil. Nous donnons la possibilité à un administrateur de modifier les concepts s'il remarque que le résultat obtenu peut être amélioré. Nous avons développé pour cela une interface qui lance des mises à jour des graphes, des liens et de la distribution en considérant la modification effectuée par l'administrateur. L'interface est présentée dans la figure 8.6 ci-dessous.

4. L'administrateur consulte les résultats de la distribution des documents

   À la fin de la distribution des documents, le système génère un graphique illustrant la nouvelle distribution des documents entre les profils. Ce graphique trace le périmètre de chaque collaborateur ainsi que les documents partagés entre

FIGURE 8.6 – Interface modification des concepts

les profils. Nous présentons dans la figure 8.7 ci-dessous un exemple d'une capture de ce graphique lors d'une distribution.

Par exemple, dans ce graphique, nous pouvons visualiser que 10 nouveaux documents ont été distribués comme suit : 7 documents ont été partagés entre tous les profils, 2 documents entre seulement les profils informaticien et ingénieur mécanique et un document a été partagé entre trois profils : l'ingénieur aéronautique, l'informaticien et le mécanicien.

La liste des profils a été créée, les graphes et les liens sémantiques ont été générés et les documents ont été distribués ; un utilisateur peut maintenant effectuer le scénario de test utilisateur. Nous détaillons les sous-étapes avec les interfaces correspondantes.

1. Un utilisateur ajoute un document pour distribution
   À travers la fenêtre présentée dans la figure 8.8, nous donnons la possibilité à un utilisateur (collaborateur) de participer à la distribution des documents de l'entreprise s'il est conscient qu'un document en particulier est important pour plusieurs autres personnes. Le document que l'utilisateur a

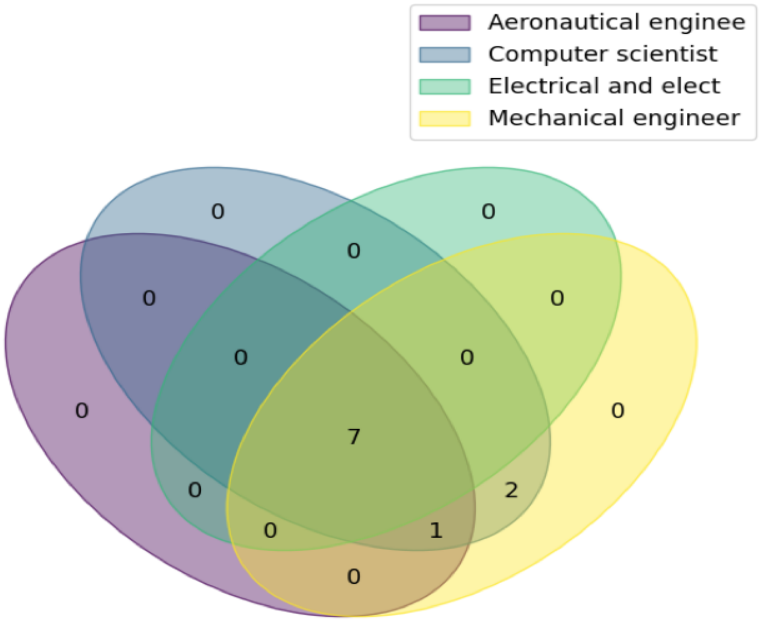


FIGURE 8.7 – Distribution des documents

ajouté comporte le contenu représenté dans la figure 8.9 suivante

À travers le bouton choisir le fichier, le fichier est chargé puis en appuyant sur le bouton indexer le système prend le relais pour :

— Analyser le document : tout d'abord le nettoyer, éliminer les articles définis, indéfinis ou partitifs ainsi que tous les séparateurs. Ensuite identifier les mots les plus fréquents, avec leurs occurrences, traiter la synonymie et l'hyperonymie pour représenter le document sous forme d'un tableau de concepts fréquents ;

— Comparer les concepts générés par rapport aux lexiques des profils avec un taux de similarité de 70% (fixé par

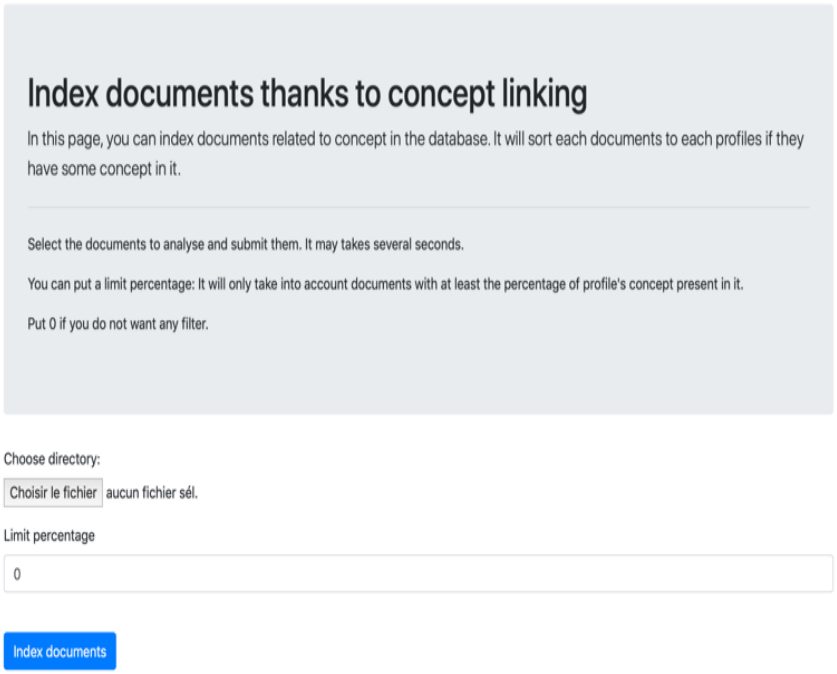


FIGURE 8.8 – Ajout des documents

```
Four-stroke-cycle engines; valve gear.—In an
engine having two rows of cylinders with
parallel crank-shafts and with common combustion chambers between the rows, each
connected with a cylinder of each row, the
inlet and exhaust valve of each combustion
chamber are actuated by a single cam on a
shaft between the cylinder rows. Fig. 1 shows
inclined cylinders D, E arranged in connection
with crank-shafts B, C and communicating
with a combustion chamber H which is nearly
spherical. The inlet and exhaust valves J 1 , K 1
are actuated by a cam L 2 through rockers M1,
push rods N, and rockers N 1 . The push rods
N of two consecutive cylinders of a row are
arranged in a single tubular housing in the
cylinder block. The cam-shaft is driven by
gearing O . . O 4 .
Valves.—The period of opening of the inlet
valve is made less than the exhaust valve,
although both actuated by the same cam, by
masking the seating of the inlet valve as shown
a t J 2 , Fig. 3. The exhaust valve seating K 2
is recessed to the same depth, but is conical
throughout, as shown in Fig. 4.
```

FIGURE 8.9 – Contenu à distribuer

l'administrateur), ce qui exige une forte similarité entre les deux représentations pour que, au final, le document

puisse être distribué au profil et indexé dans le profil correspondant ;

— Analyser les liens sémantiques : le système interprète l'existence des liens sémantiques en se basant sur le principe que nous avons détaillé dans le chapitre 7.

2. L'utilisateur consulte la liste des documents du profil.
   Un utilisateur, comme il appartient à un profil, peut consulter la liste des documents de son profil. Cette liste de document, nous la générons dans un dossier spécifique généré pour chaque profil comme représenté dans la figure 8.10 ci-dessous.

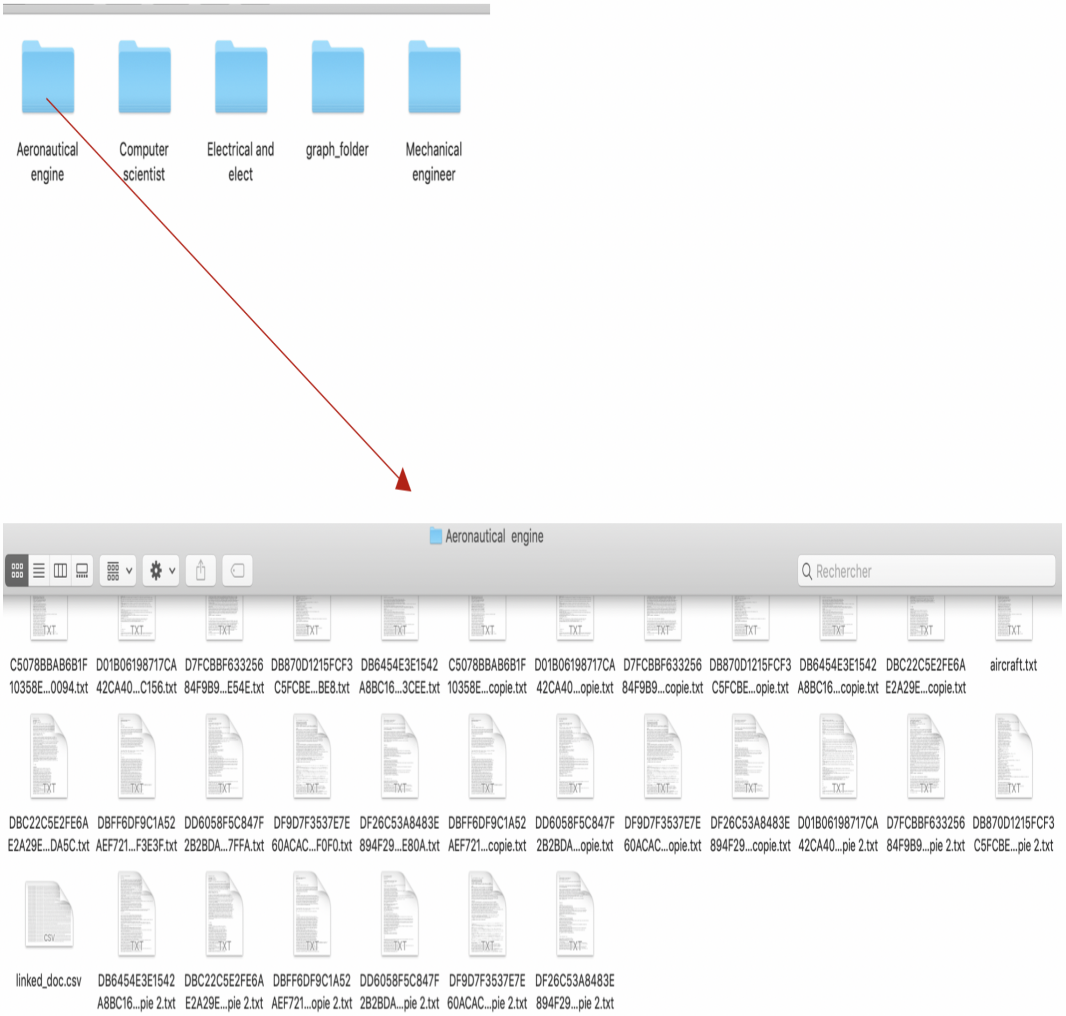


FIGURE 8.10 – Documents du profil

3. L'utilisateur voit qu'un nouveau document est ajouté à son profil :
   L'utilisateur a mis un nouveau document en distribution. Comme ce document a été ajouté récemment dans la liste des profils correspondant à travers l'infrastructure, le système envoie une notification informant que les graphes ont été modifiés (ou cas où de nouveaux concepts auraient été générés), et que le document a été indexé. Cette notification s'envoie aux utilisateurs concernés et à l'administrateur comme illustré dans la figure suivante

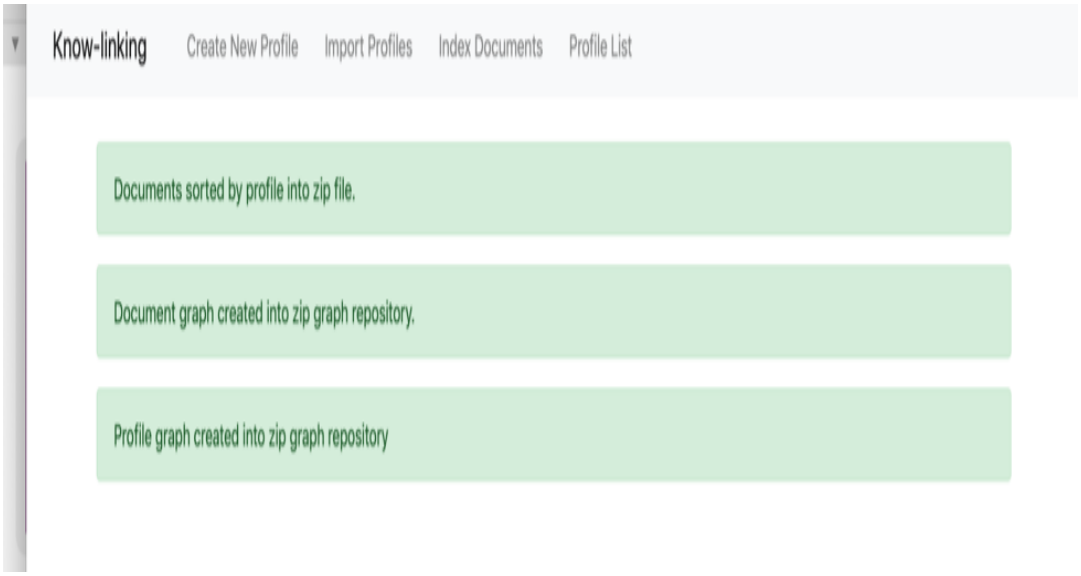


FIGURE 8.11 – Notification de mise à jour

8.8 Résultats

Nous avons effectué plusieurs tests pour vérifier sa conformité par rapport aux spécifications prédéfinies. Nous passons dans la partie suivante à l'analyse des résultats obtenus. Nous organisons la partie suivante par étape de l'approche know-linking.

8.8.1 Résultats de profilage

L'infrastructure know-linking a généré automatiquement des profils. Pour chaque profil nous avons pu récupérer les informations demandées comme : son lexique, son intitulé, ses missions et ses outils comme suit (figure 8.12 ci-dessous).

Computer scientist	develop and/or simplify algorithms, create new computing languages, determine new methods for working with computers, test new systems and designs, develop models and theories to address issues in the field, present findings to the scientific community, improve computer hardware performance, increase the efficiency of computer software and/or hardware	eclipse			computing languages, computers, UML, computer hardware, computer software
Electrical and elect	make electrical diagrams with dedicated software, make perform electronic simulations or calculations, replace defective components	None	None	None	engineer, Mission, diagrams, software, perform, simulations, calculations, prototype, tests, components, cards, equipment, faults, test, procedures, List, tools, IGEXAO, E3Series, AMASIS, CATIA, Multisim, Eagle, Source, knowledge, documents, electronics, wiki, reports, IC,
Mechanical engineer	Performing and providing technical definition dossiers for metallic andor composite structures covering design data set elaboration, Providing approval for structure design in order to propose optimized engineering technical solutions compliant with safety and airworthiness requirements, Ensuring the safe operation of aircraft with respect to structural integrity	None	None	None	Mechanical, engineer, definition, dossiers, andor, structures, design, data, elaboration, approval, structure, order, engineering, solutions, safety, airworthiness, requirements, operation, aircraft, respect, integrity, List, tools, CATIA, Source, knowledge, Design, documents, reports, tests

FIGURE 8.12 – Profils générés

Le nombre de profils générés est celui des documents de description de poste que nous avons introduits. Bien que nous ayons introduit 10 descriptions de poste, nous avons obtenu 10 profils. Nous analysons les résultats du profilage obtenus par rapport aux dimensions de l'approche know-linking que nous avons tracées lors de l'état de l'art de profilage dans le tableau 8.3 suivant.

TABLE 8.3 – Dimensions du profilage.

Filtrage hybride	Traçabilité personnalisée	Représentation des profils	Profilage automatique	Mise à jour régulière
L'algorithme de profilage basé sur le filtrage hybride est fonctionnel et nous a permis de générer automatiquement des profils.	Différents profils ont été générés nous permettant d'avoir une traçabilité personnalisée par profil.	Chaque profil suit la structure que nous avons définie : intitulé, outils, domaine, missions et lexique (liste de concepts).	L'algorithme de profilage s'exécute automatiquement lors de la mise en place de l'infrastructure avec une possibilité d'être lancé par l'administrateur.	L'administrateur lance régulièrement des requêtes de mise à jour.

8.8.2 Résultats de la génération des graphes et des liens sémantiques

Pour chaque profil nous avons obtenu une modélisation sous forme d'un graphe de connaissances. Par exemple, le graphe du profil électrique est représenté dans la figure 8.13 ci-dessous.
Les liens entre les profils sont générés dans un deuxième temps dans des fichiers CSV. Un exemple de lien est présenté dans la figure 8.14 ci-dessous. Dans ce lien nous remarquons que l'algorithme a détecté deux profils différents dans le contenu du texte. En effet, ce document parle d'un logiciel qui concerne un profil mécanicien. Généralement ce type de document est distribué aux informaticiens car son contenu tourne autour d'un logiciel, néanmoins il concerne aussi les mécaniciens.
En analysant les liens sémantiques générés nous remarquons que ces liens expriment bien un rapport entre les profils et qu'ils influencent de façon effective la distribution des documents. Cependant, un travail de filtrage doit être complété par l'analyse de contexte pour améliorer les résultats obtenus.
Après avoir généré les profils nous avons appliqué un algorithme

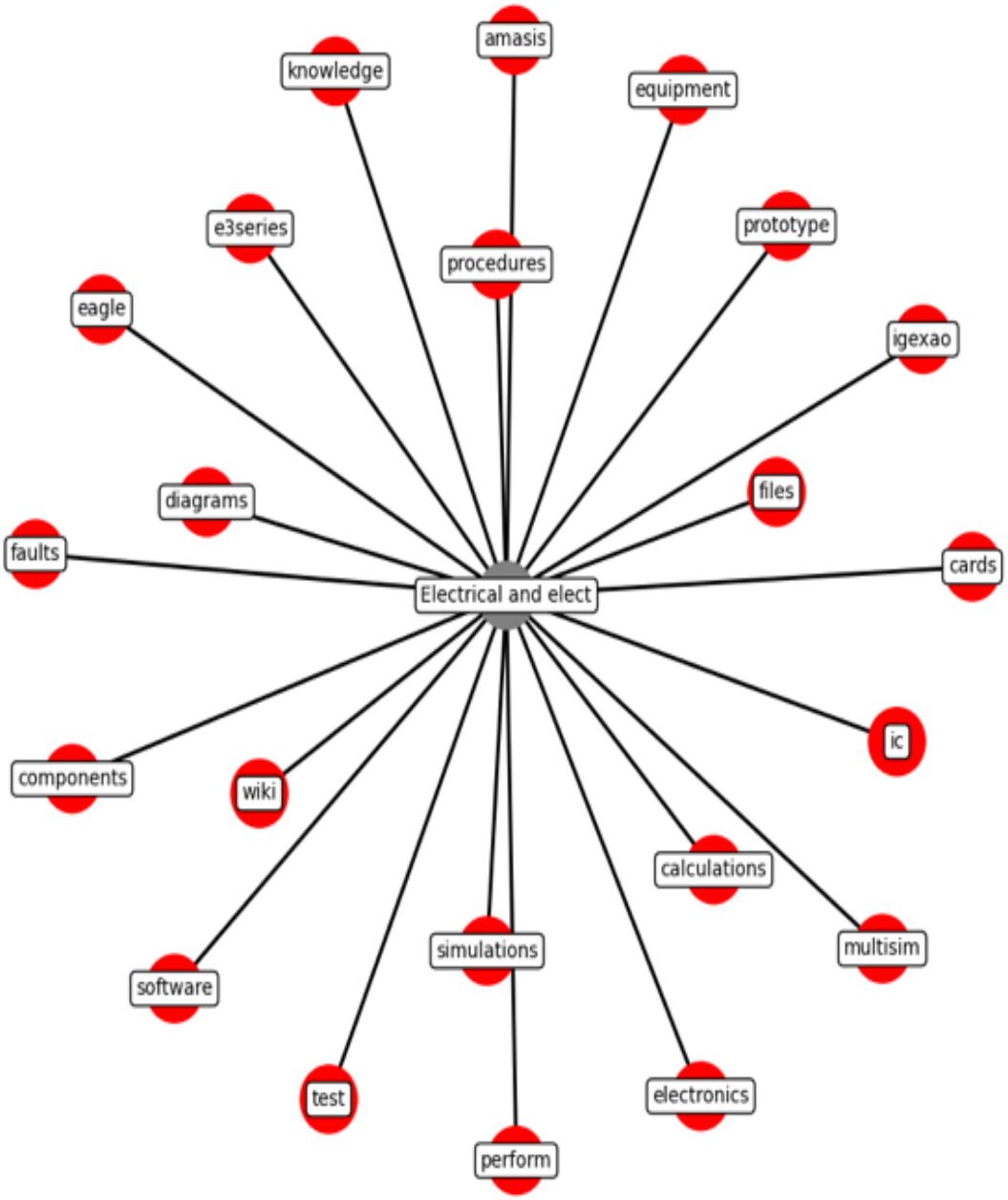


FIGURE 8.13 – Graphe du profil

d'indexation sémantique pour indexer les documents dans les profils. En parallèle un script s'est exécuté pour savoir sous quels profils les documents seront indexés.

Le script nous a généré le résultat de la distribution des documents sans considérer les liens sémantiques comme présenté dans la figure 8.15 ci-dessous.

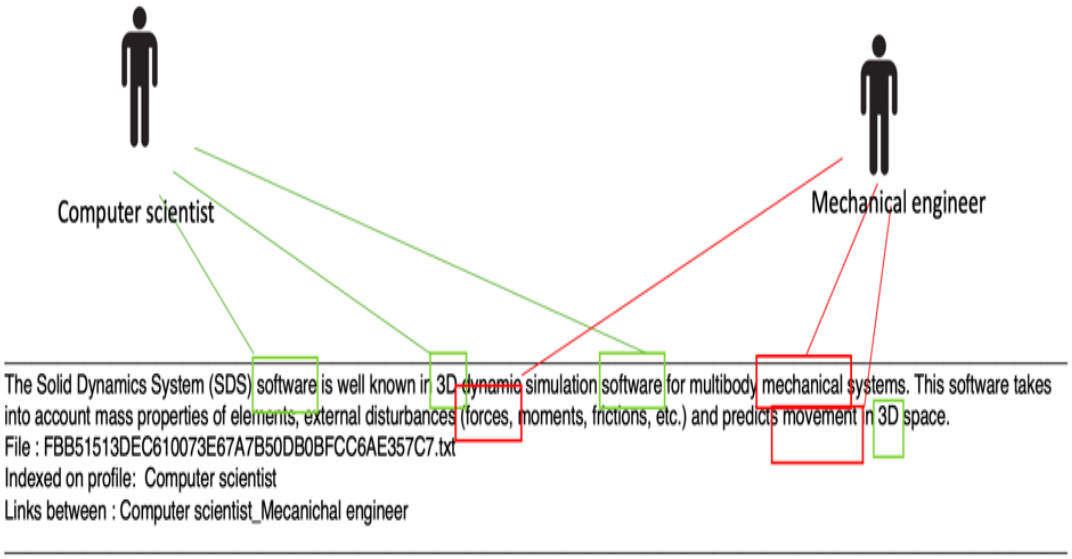


FIGURE 8.14 – Lien sémantique

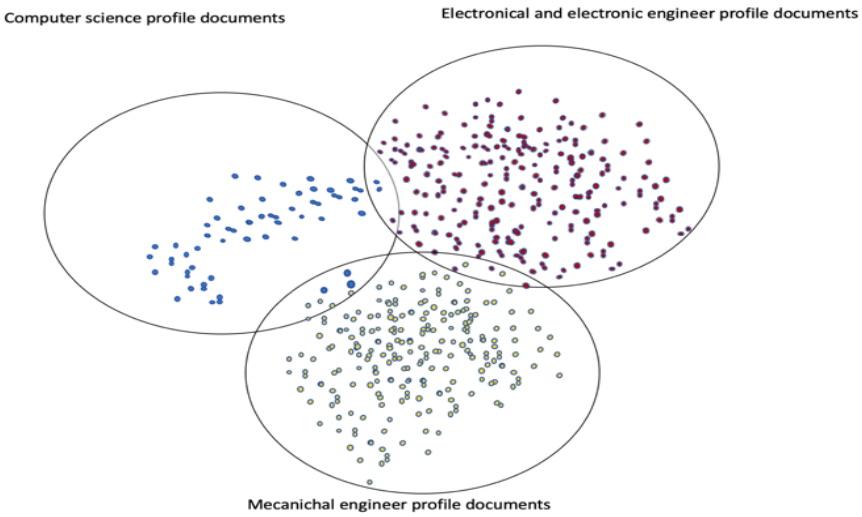


FIGURE 8.15 – Distribution des documents avant la génération des liens sémantiques

Pour comparer les résultats de la distribution obtenue en considérant les liens sémantiques nous nous basons sur la deuxième distribution générée par le même script sur les mêmes documents dans la figure 8.16 ci-dessous.

Nous constatons que les liens sémantiques découverts ont permis d'améliorer l'accès aux connaissances. Le nombre des liens découverts s'accroît en ajoutant des documents à analyser au fur et à mesure jusqu'à pouvoir analyser les 1000 documents dans le corpus.

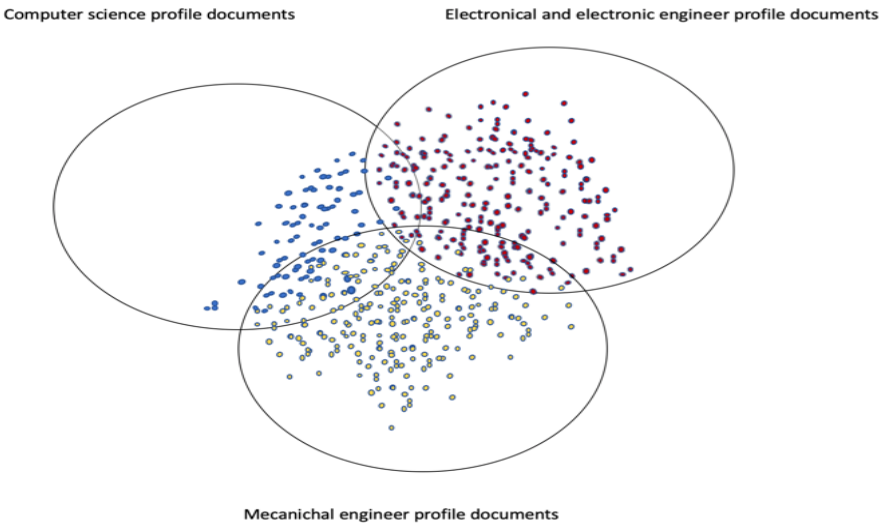


FIGURE 8.16 – Distribution des documents après la génération des liens sémantiques

Nous modélisons cette croissance, ou autrement dit le nombre des liens sémantiques vs le nombre des documents, dans le graphique suivant (figure 8.17).

8.8.3 Résultats de la distribution des documents

Pour pouvoir tester l'efficacité de nos distributions de documents nous avons appliqué deux différents algorithmes d'indexation (Bitmap et l'indexation sémantique) dans un premier temps. Puis nous avons introduit un lien vers un corpus de 1000 documents de l'aéronautique dans l'infrastructure pour que le collecteur des documents puisse les capter, les posséder automatiquement, les analyser et finalement les distribuer.
Nous avons comparé la capacité de chaque algorithme à analyser l'ensemble des documents avec des paramètres temps d'exécution et le bruit. Les résultats sont présentés dans le tableau comparatif 8.4 suivant.

Nous pouvons remarquer que Know-linking a amélioré l'accès aux documents et a réduit le bruit tout en gardant un temps d'exécution raisonnable.

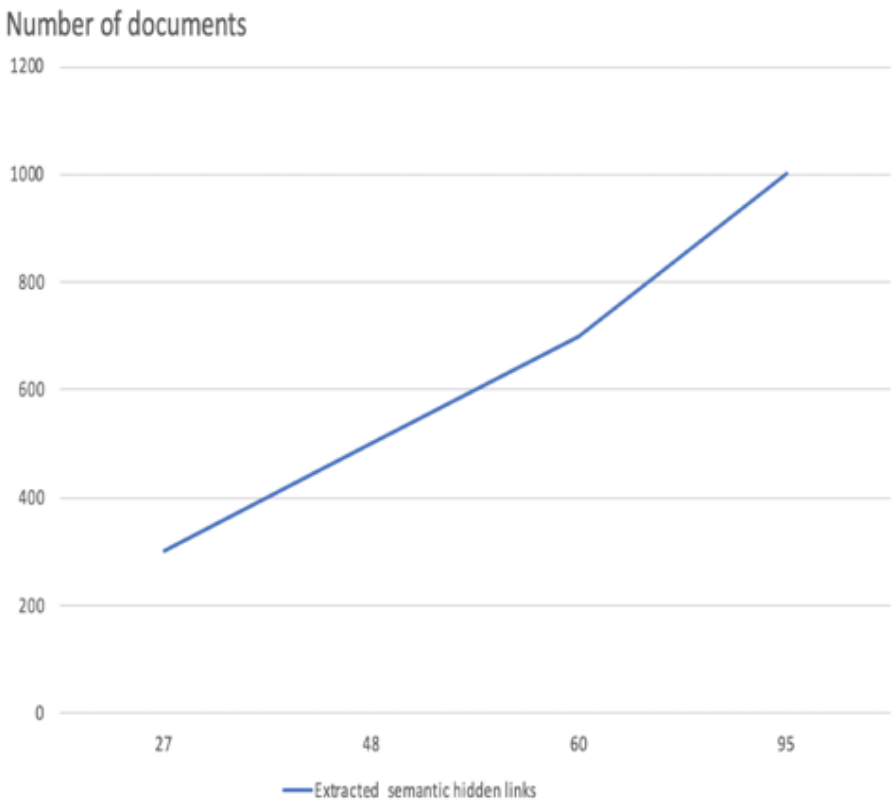


FIGURE 8.17 – Liens sémantiques générés

TABLE 8.4 – Évaluation

Approche	**Nombre de documents indexé**	**Bruit**	**Temps moyen pour retrouver le document**
knowlinking	1000	12	6mins
Semantic indexing	1000	68	6mins
Bitmap indexing	1000	120	14mins

Pour plus d'analyses de la distribution des documents, nous prenons l'exemple du contenu dans la figure ci-dessous 8.18.
Théoriquement, ce document ne concerne que les collaborateurs ayant le profil «électrique». Par contre, lorsque l'on met ce document en analyse avec know-linking, nous obtenons la distribution

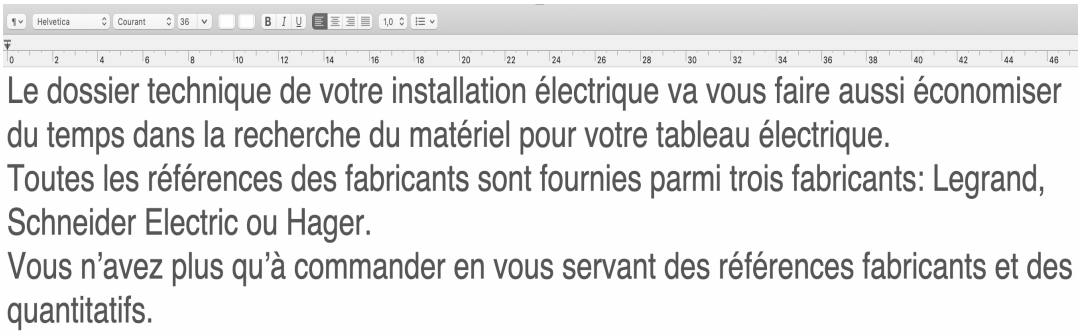


FIGURE 8.18 – Contenu à distribuer

suivante (dans la figure 8.19)

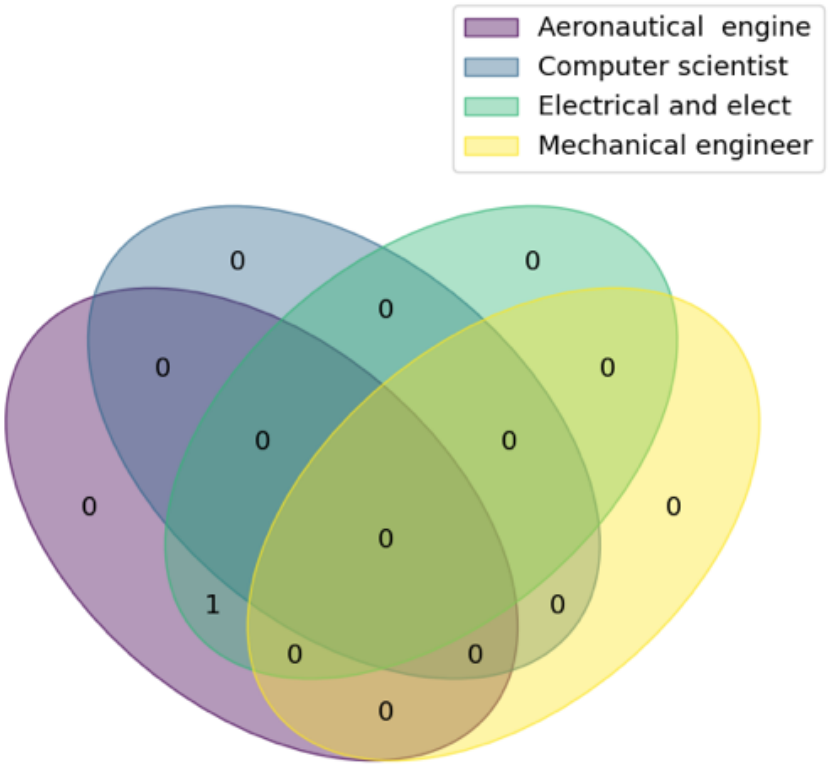


FIGURE 8.19 – Distribution du contenu

Ce document doit être partagé entre les deux profils électrique et aéronautique. Si on revient au graphe de profil aéronautique, nous remarquons l'existence de concepts liés à l'électrique déjà extraits depuis son contenu, ce qui explique le besoin de ce profil d'accéder à ce document.

8.9 Conclusion

La mise en place d'une infrastructure (framework) logicielle pour tester l'application de Know-linking nous a démontré la possibilité d'avoir une solution d'entreprise qui permet de résoudre les problématiques que nous avons dégagées, et de considérer à la fois toutes les dimensions que nous avons émises lors de l'étude de l'état de l'art. Cependant, ces résultats peuvent être encore améliorés en ajoutant de nouvelles perspectives sur l'infrastructure de tests telles que :

— Intégrer des plugins dans l'environnement technique du collaborateur pour détecter automatiquement l'ajout ou la création d'un nouveau document, puis les analyser directement.

— Un module de gestion d'accès des identités (IAM) qui doit compléter l'infrastructure pour générer les accès aux documents en considérant les droits d'accès des profils aux ressources. Plus précisément, certains documents doivent restés privés ou confidentiels même s'ils intéressent d'autres profils. Avec la gestion des accès et des identités il est possible de respecter la politique de distribution des accès dans l'entreprise.

Mis à part ces améliorations que nous devons adopter pour l'infrastructure de tests, d'autres réflexions d'amélioration et des perspectives de recherche au niveau de l'approche seront détaillées dans la dernière partie du rapport.

Quatrième partie

Conclusion et perspectives

Chapitre 9

Conclusion et perspectives

*"Une conclusion, c'est quand vous en avez assez de penser."*Herbert Albert Fisher

Sommaire

9.1 Conclusion

La définition tripartite antonienne voit la connaissance comme une « croyance vraie justifiée ». Des travaux ultérieurs de E.Gettier en 1923 montrent que pour définir la connaissance il faut aller plus en profondeur et s'articuler sur d'autres conditions philosophiques dites « internalistes » et « externalistes » pour renforcer les trois conditions (croyance, vraie et justifiée). Ce renforcement repose sur l'univers social de la connaissance. Que ce soit les croyances ou les connaissances, elles sont largement partagées au sein d'une communauté, acquises et transmises au sein d'un univers social [Tiercelin 2012].

D'autres études s'intéressent à l'évocation de cette relation de la connaissance avec l'environnement social. Par exemple [Johannessen *et al.* 2002] posent la question **«comment les faits sociaux sont transformés en données, les données en informations et les informations en connaissances ?»**.

L'influence de la connaissance sur son environnement social est bi-directionnelle dans le sens où cet environnement est influencé par la connaissance, et en même temps celle-ci a un impact sur l'environnement.

Pour expliquer cette influence [Johannessen *et al.* 2002] ont présenté un modèle de transformation qui met en évidence l'importance de l'environnement social dans un cycle de transformation de la donnée en information, l'information en connaissance et la connaissance en fait social. Le modèle présenté est inspiré d'un cercle fermé qui n'a pas de début ni de fin. Ce modèle, qui commence par les faits sociaux, s'achève aussi par ces faits. Les auteurs ont présenté ce modèle conceptuel dans l'illustration 9.1 suivante. Ce modèle conceptuel souligne que la connaissance est incluse dans

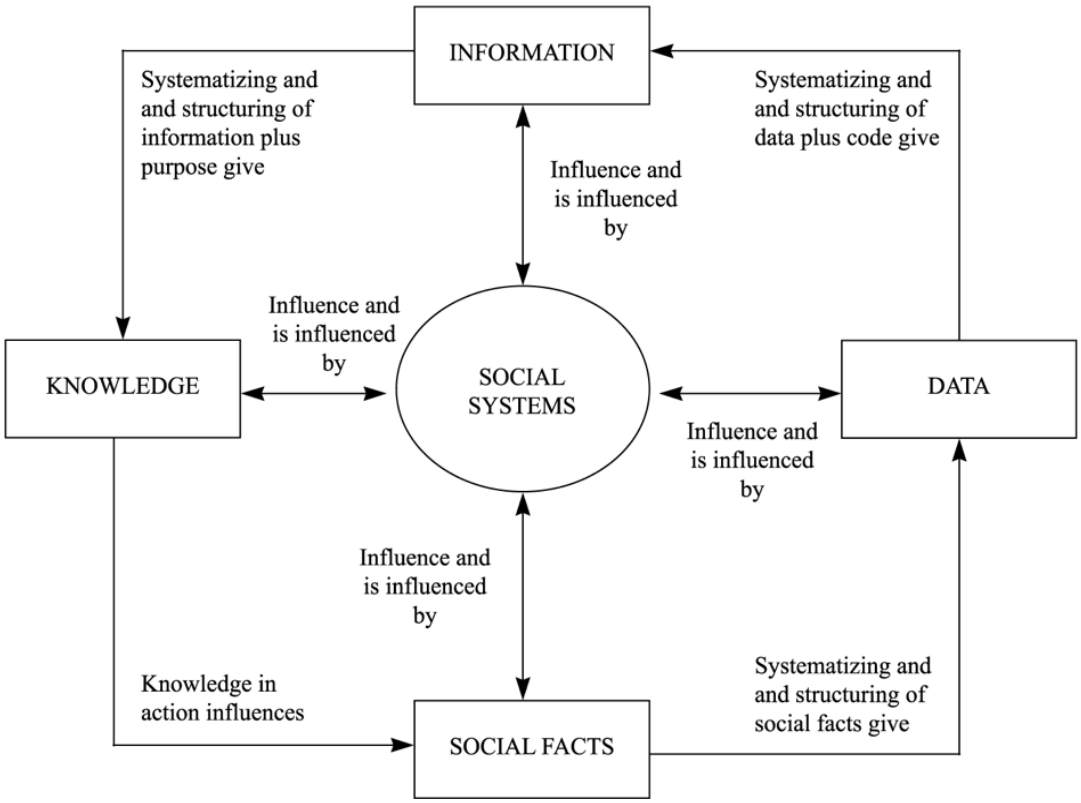


FIGURE 9.1 – Modèle conceptuel [Johannessen *et al.* 2002]

un cycle en échange permanent avec l'univers social. Ce système social est en fait composé de quatre sous-systèmes dans lesquels, pour chacun, il existe un mode d'expression spécifique, notamment : le sous-système culturel, le sous-système coopératif, le sous-système politique et le sous-système économique. Ainsi la connaissance est influencée par l'ensemble de ces sous-systèmes et elle les influence parallèlement.

Dans ce contexte, nous mettons en valeur l'influence que la connaissance peut avoir sur les différents composants du système social, notamment sur l'organisation à travers ses acteurs et ses opérations. Nos travaux de recherche répondent aux enjeux du partage de la connaissance dans un environnement complexe, une grande entreprise essentiellement caractérisée par des processus multidisciplinaires, d'un nombre de collaborateurs assez important et de projets de longue durée. Cet environnement, Nonaka [Nonaka &

Konno 1998] le considère comme concept philosophique et l'appelle «BA», un environnement où la connaissance est créée, transformée et ainsi partagée. L'observation de ce contexte nous a permis de dégager des questions de recherche. Après des études bibliographiques et des travaux de tests nous avons trouvé quelques réponses. Chacune des réponses présente un facteur clef de la construction de l'approche proposée, d'une dimension ou d'un chemin que nous avons suivi.

La première question que nous avons posée était : **Comment générer la connaissance partagée entre les différents acteurs d'une façon régulière et dynamique ?**

La structuration des besoins des acteurs de l'entreprise sous forme de profils de collaborateurs nous a permis de tracer les périmètres des besoins en termes de connaissances de chacun dans l'entreprise. Pour chaque profil nous avons procédé à une analyse organisationnelle qui nous a permis de déterminer son affiliation, son domaine, ses missions et l'équipe avec laquelle il travaille. Nous avons ainsi identifié l'influence organisationnelle de la connaissance. Une autre analyse opérationnelle permet de déterminer les tâches et les projets dans lesquels le contributeur est impliqué. Nous avons alors étendu notre étude à la dimension opérationnelle d'une entreprise. Les méthodes de traitement du langage naturel, et plus particulièrement le «bag-of concepts», nous a permis de définir pour chaque profil une liste de concepts représentant son glossaire du domaine. Les profils que nous avons construits ont permis aux collaborateurs de partager les connaissances selon deux axes : un axe temporel qui considère les collaborateurs n'appartenant pas à l'entreprise pendant la même période de temps, et un autre axe où les collaborateurs travaillent dans la même entreprise mais dans des services ou des domaines différents. À travers la modélisation des graphes de connaissances de chaque profil, nous avons pu établir des liens

entre chaque profil et les connaissances qu'il détient. Les liens sémantiques que nous avons découverts à l'aide d'un algorithme basé sur les relations entre les concepts dans les documents produits par chaque contributeur ont complété ce travail par la découverte d'autres interactions entre les profils qui ont une influence sur le partage des connaissances. Par rapport à la bag-of, concept des profils, nous avons proposé d'alimenter cette représentation des concepts d'une façon automatique (après chaque analyse des nouveaux documents du profil) à l'aide de méthodes de l'analyse du contenu. Cet enrichissement continuel de la représentation du profil, par de nouveaux concepts, engendre par la suite une mise à jour automatique de la génération des liens et de la distribution des documents. Nous répondons de cette manière aux enjeux du dynamisme d'une approche de génération de connaissances partagées.

La seconde question que nous avons posée concernait l'efficacité d'une approche de traçabilité de connaissances, **« Comment assurer une traçabilité efficace des connaissances ? »**

Le profilage des collaborateurs assure une dimension de personnalisation pour la traçabilité des connaissances. C'est à travers les graphes de connaissances que nous avons assuré des liens entre les profils et les connaissances. Les connaissances sont donc désormais liées aux profils qui leur correspondent et par conséquent les traces de ces connaissances le sont aussi. Ce lien assure un suivi des traces des connaissances de chaque profil. Une approche de traçabilité efficace est une approche qui est bien ciblée !

La troisième question de recherche que nous avons posée dans la première partie de ce manuscrit de thèse était : **Comment réduire le bruit ?**

Le bruit est lié à des connaissances non intéressantes pour les collaborateurs. Le travail de personnalisation que nous avons effectué

pour obtenir une approche ciblée de génération de connaissances partagées assure le partage de la connaissance selon deux principes différents : premièrement, un document est partagé à un profil donné suite à une comparaison de similarité entre les concepts d'un document et ceux du profil. Deuxièmement, un document peut être partagé avec un autre profil car il exprime un lien sémantique entre deux profils ou plus. Ce travail a renforcé les liens entre les collaborateurs et les documents qui les intéressent et, par conséquent, a réduit d'une manière considérable le bruit (d'après ce que nos résultats ont prouvé avec les tests que nous avons effectués sur deux domaines, l'aéronautique et dans le secteur de l'énérgie et l'environnement). L'apport de Know-linking est alors d'assurer un espace de génération de connaissances partagées ciblé et enrichi d'une manière dynamique. Cet espace contribue à la relation d'influence mutuelle de connaissances, environnement qui est dans notre cas l'organisation. Des pistes d'amélioration se sont ouvertes à l'issue des travaux de recherche et des développements que nous avons menés dans le cadre de cette thèse. Nous les détaillons dans la section suivante intitulée limites et perspectives.

9.2 Limites et perspectives

À l'issue de la phase de tests et d'analyse des résultats, nous avons détecté certaines limites dans l'approche Know-linking. Ces limites ouvrent des perspectives pour continuer le travail sur le framework de test ou sur l'approche en elle-même. Concernant l'infrastructure Know-linking, nous citons les limites/perspectives que nous avons identifiées comme suit :

— Améliorer la semi-supervision : bien que les résultats des algorithmes semi-supervisés que nous avons proposés soient prometteurs, la semi-supervision peut encore être améliorée

par l'utilisation d'un jeu de données (des données labellisées) de taille encore plus importante que celle utilisée dans nos tests. Ces données peuvent être des documents ou des bases de données de ressources humaines plus volumineuses.

— Gérer les accès : dans certaines entreprises, les accès aux documents sont gérés selon une organisation spécifique où chaque collaborateur a le droit d'accéder aux documents en respectant des niveaux de confidentialité dans l'entreprise. Par exemple, supposons que dans une entreprise le niveau de confidentialité 1 soit accordé à tout le monde (des stagiaires, des prestataires, à tous les collaborateurs) et que par contre le niveau 2 soit accordé uniquement aux collaborateurs permanents, le niveau est spécifique pour les chefs de projets d'entreprise . L'infrastructure ou le framework de test que nous avons développé ne prend pas en compte, dans sa version actuelle, la distribution des documents selon la politique de gestion d'accès de l'entreprise. Nous envisageons dans un travail futur d'intégrer un autre composant logiciel qui soit lié à l'annuaire de l'entreprise pour assurer la gestion des accès et des identités (IAM).

— Améliorer la capture de nouveaux documents : dans la version actuelle de test de l'infrastructure Know-linking, nous proposons un espace de collecte de nouveaux documents à analyser lié à une base de documents à partir d'une seule plateforme de travail. Cette option nous a procuré un environnement de test de l'approche. Cependant, elle représente une limite face à la capacité de l'approche de capter et d'analyser un maximum de nouveaux documents. Nous proposons donc d'améliorer la collecte de nouveaux documents en développant des plugins qui servent à capter les nouveaux documents de travail crées par les collaborateurs. Ces plugins

doivent être intégrés dans différents environnements de travail ou dans différents systèmes d'information. En analysant les résultats des tests obtenus, nous avons eu certaines réflexions qui peuvent compléter notre approche Know-linking. Par exemple :

— Détecter les verbes dans les liens sémantiques et les analyser : nous avons commencé en premier lieu un travail de détection des verbes avec l'analyse morphosyntaxique. Ce premier pas peut être complété par l'analyse de ces verbes pour comprendre la nature du lien exprimé par le verbe, ce qui permettrait d'enrichir les graphes de connaissances et également d'avoir plus de précisions au niveau des liens sémantiques générés.

— Analyse de contexte : en complément de la perspective précédente, l'analyse des verbes peut mener à la compréhension du contexte de la connaissance. En déterminant la sémantique des verbes nous pouvons contextualiser les connaissances et les traces des connaissances. Comme exemple de contexte nous citons : la justification des choix .

— La recommandation de nouvelles connaissances : la recommandation [Duthil 2012] des connaissances peut compléter notre travail par un mécanisme qui s'intégrerait dans l'environnement technique des collaborateurs en leur proposant la bonne connaissance au bon moment !

Bibliographie

[A & Wáng 2017] Thomas A et Yì N. Wáng. *Resolving distributed knowledge.* Artificial Intelligence, vol. 252, pages 1–21, 2017.

[Abderrahim *et al.* 2020] Elamin Abderrahim, Nada Matta et Hassan Atifi. *Know-Linking : An approach to generate shared knowledge between several actors in companies application in aerospace manufacturer.* In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 1218–1223. IEEE, 2020.

[Abel *et al.* 2002] Marie-Hélène Abel, Dominique Lenne et Omar Cissé. *E-learning et Web sémantique : Le projet MEMORAe.* Actes Électroniques des Journées Scientifiques Web Sémantique, 2002.

[Ahmed Awan *et al.* 2019] Malik Nabeel Ahmed Awan, Sharifullah Khan, Khalid Latif et Asad Masood Khattak. *A New Approach to Information Extraction in User-Centric E-Recruitment Systems.* Applied Sciences, vol. 9, no. 14, page 2852, 2019.

[Arbi 2014] Zied Arbi. *Natural Language Processing.* 04 2014.

[Baclic *et al.* 2020] Oliver Baclic, Matthew Tunis, Kelsey Young, Coraline Doan, Howard Swerdfeger et Justin Schonfeld. *Le traitement du langage naturel (TLN), une sous-zone d'intelligence artificielle.* 2020.

[Bekhti 2003] Smain Bekhti. *DypKM : Un processus dynamique de définition et de réutilisation de mémoires de projets.* 2003.

[Berners-Lee *et al.* 2001] Tim Berners-Lee, James Hendler et Ora Lassila. *The semantic web.* Scientific american, vol. 284, no. 5, pages 34–43, 2001.

[Bessire & Mesure 2009] Dominique Bessire et Hervé Mesure. *Penser l'entreprise comme communauté : fondements, définition et implications.* Management Avenir, no. 10, pages 30–50, 2009.

[Bestgen 2006] Yves Bestgen. *Improving text segmentation using latent semantic analysis : A reanalysis of choi, wiemer-hastings, and moore (2001).* Computational linguistics, vol. 32, no. 1, pages 5–12, 2006.

[Bijalwan *et al.* 2014] Vishwanath Bijalwan, Vinay Kumar, Pinki Kumari et Jordan Pascual. *KNN based machine learning approach for text and document mining.* International Journal of Database Theory and Application, vol. 7, no. 1, pages 61–70, 2014.

[Bird 2006] Steven Bird. *NLTK : the natural language toolkit.* In Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions, pages 69–72, 2006.

[Bissay *et al.* 2008] Aurélie Bissay, Philippe Pernelle, Arnaud Lefebvre et Abdelaziz Bouras. *Démarche d'intégration des connaissances au système PLM.* In MOSIM'08, page 8, 2008.

[Bond] Thomas Bond. *Le profilage criminel.*

[Bouhriz *et al.* 2014] Nadia Bouhriz, Faouzia Benabbou et EL Habib Benlahmar. *Approche hybride d'indexation conceptuelle des documents en Arabe.* 05 2014.

[Brill 1992] Eric Brill. *A simple rule-based part of speech tagger.* Rapport technique, PENNSYLVANIA UNIV PHILADELPHIA DEPT OF COMPUTER AND INFORMATION SCIENCE, 1992.

[Cambria] Erik Cambria. Sentic computing [electronic resource] : techniques, tools, and applications. Springer.

[Cambria & White 2014] Erik Cambria et Bebo White. *Jumping NLP curves : A review of natural language processing research.* IEEE Computational intelligence magazine, vol. 9, no. 2, pages 48–57, 2014.

[Champin *et al.* 2004] Pierre-Antoine Champin, Yannick Prié et Alain Mille. *MUSETTE : a framework for Knowledge from Experience.* In Extraction et gestion des connaissances (EGC'2004)(article court), pages 129–134, 2004.

[Chan & Ioannidis 1998] Chee-Yong Chan et Yannis E Ioannidis. *Bitmap index design and evaluation.* In Proceedings of the 1998 ACM SIGMOD international conference on Management of data, pages 355–366, 1998.

[Chandler 2007] Daniel Chandler. Semiotics : the basics. Routledge, 2007.

[Chatzopoulou *et al.* 2011] Gloria Chatzopoulou, Magdalini Eirinaki, Suju Koshy, Sarika Mittal, Neoklis Polyzotis et Jothi Swarubini Vindhiya Varman. *The QueRIE system for Personalized Query Recommendations.* IEEE Data Eng. Bull., vol. 34, no. 2, pages 55–60, 2011.

[Chen *et al.* 2010] Jilin Chen, Rowan Nairn, Les Nelson, Michael Bernstein et Ed Chi. *Short and tweet : experiments on recommending content from information streams.* In Proceedings of the SIGCHI conference on human factors in computing systems, pages 1185–1194, 2010.

[Cheng *et al.* 2011] James Cheng, Yiping Ke, Ada Wai-Chee Fu et Jeffrey Xu Yu. *Fast graph query processing with a low-cost index.* The VLDB Journal, vol. 20, no. 4, pages 521–539, 2011.

[Choi & Han 2008] Okkyung Choi et Sang Yong Han. *Personalization of rule-based web services.* Sensors, vol. 8, no. 4, pages 2424–2435, 2008.

[Chowdhury 2003] Gobinda G Chowdhury. *Natural language processing.* Annual review of information science and technology, vol. 37, no. 1, pages 51–89, 2003.

[Clarke 1993] Roger Clarke. *Profiling : A hidden challenge to the regulation of data surveillance.* JL & Inf. Sci., vol. 4, page 403, 1993.

[CoE] *CoE2010 Recommendation CM/Rec(2010)13 of the Committee of Ministers to member states on the protection of individuals with regard to automatic processing of personal data in the context of profiling. Adopted by the Committee of Ministers on 23 November 2010 at the 1099th meeting of the Ministers' Deputies, howpublished =* `https://www.coe.int/en/web/data-protection/home`*, note = Accessed : 2021-05-05.*

[Corbel 1997] JC Corbel. *Méthodologie de retour d'expérience : démarche MEREX de Renault.* Hermès, vol. 129, 1997.

[Cori 2008] Marcel Cori. *Des méthodes de traitement automatique aux linguistiques fondées sur les corpus.* Langages, no. 3, pages 95–110, 2008.

[Cufoglu 2014] Ayse Cufoglu. *User profiling-a short review.* International Journal of Computer Applications, vol. 108, no. 3, 2014.

[Dalkir 2013] Kimiz Dalkir. Knowledge management in theory and practice. Routledge, 2013.

[Dalloux 2020] Clément Dalloux. *Fouille de texte et extraction d'informations dans les données cliniques*. PhD thesis, Université de Rennes 1, 2020.

[Damien 2012] T Damien. *Marc Bloch, Apologie pour l'histoire ou Métier d'historien, Extrait : commentaire*. Publications Pimido, 2012.

[Deerwester *et al.* 1990] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer et Richard Harshman. *Indexing by latent semantic analysis*. Journal of the American society for information science, vol. 41, no. 6, pages 391–407, 1990.

[Delalleau & Larochelle 2007] Olivier Delalleau et Hugo Larochelle. *Algorithme des k plus proches voisins*, 2007.

[Demonet-Launay 1994] Marie-Luce Demonet-Launay. *«Si les signes vous fâchent...», inférence naturelle et science des signes à la Renaissance*. Réforme, Humanisme, Renaissance, vol. 38, no. 1, pages 7–44, 1994.

[Di Jorio *et al.* 2007] Lisa Di Jorio, Céline Fiot, Lylia Abrouk, Danièle Hérin et Maguelonne Teisseire. *Enrichissement d'ontologie : Quand les motifs séquentiels labellisent des relations*. In BDA : Bases de Données Avancées, 2007.

[Dieng-Kuntz *et al.* 2006] Rose Dieng-Kuntz, David Minier, Marek Ržička, Frédéric Corby, Olivier Corby et Laurent Alamarguy. *Building and using a medical ontology for knowledge management and cooperative work in a health care network*. Computers in Biology and Medicine, vol. 36, no. 7-8, pages 871–892, 2006.

[Domingos & Pazzani 2004] Pedro M. Domingos et Michael J. Pazzani. *On the Optimality of the Simple Bayesian Classifier under Zero-One Loss.* Machine Learning, vol. 29, pages 103–130, 2004.

[Dumal *et al.* 2017] PAA Dumal, WKD Shanika, SAD Pathinayake et Thanuja Chandani Sandanayake. *Adaptive and automated online assessment evaluation system.* In 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pages 1–8. IEEE, 2017.

[Duthil 2012] Benjamin Duthil. *De l'extraction des connaissances à la recommandation.* PhD thesis, Ecole Nationale Supérieure des Mines d'Alès, 2012.

[Eco 1979] Umberto Eco. A theory of semiotics, volume 217. Indiana University Press, 1979.

[Eke *et al.* 2019] Christopher Ifeanyi Eke, Azah Anir Norman, Liyana Shuib et Henry Friday Nweke. *A survey of user profiling : State-of-the-art, challenges, and solutions.* IEEE Access, vol. 7, pages 144907–144924, 2019.

[Ermine 2003] Jean-Louis Ermine. La gestion des connaissances. Hermès sciences publications, 2003.

[Fayard 2002] Pierre Fayard. Le concept de" ba" dans la voie japonaise de la création du savoir. Ambassade de France à Tokyo, Service pour la science et la technologie, 2002.

[Feldman & Dagan 1995] Ronen Feldman et Ido Dagan. *Knowledge Discovery in Textual Databases (KDT).* In KDD, volume 95, pages 112–117, 1995.

[Ferraris *et al.* 2013] Valeria Ferraris, Francesca Bosco, G Cafiero, Elena D'Angelo et Y Suloyeva. *Defining profiling.* Available at SSRN 2366564, 2013.

[Forman 2003] George Forman. *An extensive empirical study of feature selection metrics for text classification [J].* Journal of Machine Learning Research - JMLR, vol. 3, 03 2003.

[Gani *et al.* 2016] Abdullah Gani, Aisha Siddiqa, Shahaboddin Shamshirband et Fariza Hanum. *A survey on indexing techniques for big data : taxonomy and performance evaluation.* Knowledge and information systems, vol. 46, no. 2, pages 241–284, 2016.

[Geetha *et al.* 2018] G Geetha, M Safa, C Fancy et D Saranya. *A hybrid approach using collaborative filtering and content based filtering for recommender system.* In Journal of Physics : Conference Series, volume 1000, page 012101. IOP Publishing, 2018.

[Ghania & Nora 2018] Oularbi Ghania et Ould Fella Nora. *Réalisation d'une platforme de centralisation et de supervision des logs générés par les équipements réseaux de l'UMMTO.* PhD thesis, Université Mouloud Mammeri, 2018.

[Ghosh *et al.* 2020] Soumadip Ghosh, Arnab Hazra et Abhishek Raj. *A Comparative Study of Different Classification Techniques for Sentiment Analysis.* International Journal of Synthetic Emotions (IJSE), vol. 11, no. 1, pages 49–57, 2020.

[Gilloux 1989] Michel Gilloux. *Traitement automatique des langues naturelles.* In Annales des télécommunications, volume 44, pages 301–316. Springer, 1989.

[Ginzburg 1989] Carlo Ginzburg. *Traces. Racines d'un paradigme indiciaire.* Mythes, emblèmes, traces. Morphologie et histoire, pages 139–180, 1989.

[Godoy & Amandi 2005] Daniela Godoy et Analia Amandi. *User profiling in personal information agents : a survey.* The

Knowledge Engineering Review, vol. 20, no. 4, pages 329–361, 2005.

[Grčar *et al.* 2005] Miha Grčar, Dunja Mladenič et Marko Grobelnik. *User profiling for interest-focused browsing history.* In Proceedings of the Workshop on End User Aspects of the Semantic Web, Ljubljana, Slovenia, 2005.

[Grimnes 2003] Gunnar Astrand Grimnes. *Learning knowledge rich user models from the Semantic Web.* In International Conference on User Modeling, pages 414–416. Springer, 2003.

[Grundstein 2009] Michel Grundstein. *GAMETH® : a constructivist and learning approach to identify and locate crucial knowledge.* International Journal of Knowledge and Learning, vol. 5, no. 3-4, pages 289–305, 2009.

[Guan & Zhou 2002] Jihong Guan et Shuigeng Zhou. *Pruning training corpus to speedup text classification.* In International Conference on Database and Expert Systems Applications, pages 831–840. Springer, 2002.

[Guarino & Giaretta 1995] Nicola Guarino et Pierdaniele Giaretta. *Ontologies and knowledge bases.* Towards very large knowledge bases, pages 1–2, 1995.

[Guillén *et al.* 2017] Antonio Guillén, Yoan Gutiérrez et Rafael Muñoz. *Natural Language Processing Technologies for Document Profiling.* In RANLP, pages 284–290, 2017.

[Gündem & Armağan 2006] Tİ Gündem et Ö Armağan. *Efficient storage of healthcare data in XML-based smart cards.* Computer methods and programs in biomedicine, vol. 81, no. 1, pages 26–40, 2006.

[Han *et al.* 2013] Lixin Han, Guihai Chen et Ming Li. *A method for the acquisition of ontology-based user profiles.* Advances in Engineering Software, vol. 65, pages 132–137, 2013.

[Harish & Revanasiddappa 2017] BS Harish et MB Revanasiddappa. *A comprehensive survey on various feature selection methods to categorize text documents.* International Journal of Computer Applications, vol. 164, no. 8, pages 1–7, 2017.

[Harris 1954] Zellig S Harris. *Distributional structure.* Word, vol. 10, no. 2-3, pages 146–162, 1954.

[Harter 1986] Stephen P Harter. Online information retrieval : Concepts, principles, and techniques. Academic Press Professional, Inc., 1986.

[Harvey 1973] Gideon Harvey. *Archelogia Philosophica Nova or, New Principles of Philosophy, Containing : Philosophy in General ; Metaphysicks, or Ontology ; Dynamilogy, or a Discourse of Power ; Religio Philosophi, or Natural Theology ; Physicks, or Natural Philosophy.* 1973.

[Henry & Moscovici 1968] Paul Henry et Serge Moscovici. *Problèmes de l'analyse de contenu.* Langages, no. 11, pages 36–60, 1968.

[Hildebrandt & Gutwirth 2008] Mireille Hildebrandt et Serge Gutwirth. Profiling the european citizen. Springer, 2008.

[Hildebrandt 2009] Mireille Hildebrandt. *Who is profiling who ? Invisible visibility.* In Reinventing data protection ?, pages 239–252. Springer, 2009.

[Hofmann 1999] Thomas Hofmann. *Probabilistic latent semantic indexing.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57, 1999.

[Holmes & Holmes 2008] Ronald M Holmes et Stephen T Holmes. Profiling violent crimes : An investigative tool. Sage, 2008.

[Hotho *et al.* 2005] Andreas Hotho, Andreas Nürnberger et Gerhard Paaß. *A brief survey of text mining.* In Ldv Forum, volume 20, pages 19–62. Citeseer, 2005.

[Huang *et al.* 2012] Zhengxing Huang, Xudong Lu, Huilong Duan et Chenhui Zhao. *Collaboration-based medical knowledge recommendation.* Artificial intelligence in medicine, vol. 55, no. 1, pages 13–24, 2012.

[Hüllermeier *et al.* 2008] Eyke Hüllermeier, Johannes Fürnkranz, Weiwei Cheng et Klaus Brinker. *Label ranking by learning pairwise preferences.* Artificial Intelligence, vol. 172, no. 16-17, pages 1897–1916, 2008.

[Ikonomakis *et al.* 2005] M Ikonomakis, Sotiris Kotsiantis et V Tampakas. *Text classification using machine learning techniques.* WSEAS transactions on computers, vol. 4, no. 8, pages 966–974, 2005.

[Ismail *et al.* 2018] Mohammed Ismail, K Bhanu Prakash et M Nagabhushana Rao. *Collaborative filtering-based recommendation of online social voting.* International Journal of Engineering and Technology (UAE), vol. 7, no. 3, pages 1504–1507, 2018.

[Jain *et al.* 2020] Shubham Jain, Amy de Buitléir et Enda Fallon. *A review of unstructured data analysis and parsing methods.* In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), pages 164–169. IEEE, 2020.

[Jettakul *et al.* 2018] Amarin Jettakul, Chavisa Thamjarat, Kawin Liaowongphuthorn, Can Udomcharoenchaikit, Peerapon Vateekul et Prachya Boonkwan. *A comparative study on various deep learning techniques for Thai NLP lexical*

and syntactic tasks on noisy data. In 2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE), pages 1–6. IEEE, 2018.

[Jirapanthong & Zisman 2009] Waraporn Jirapanthong et Andrea Zisman. *Xtraque : traceability for product line systems.* Software & Systems Modeling, vol. 8, no. 1, pages 117–144, 2009.

[Johannessen *et al.* 2002] Jon-Arild Johannessen, Johan Olaisen et Bjørn Olsen. *Aspects of a systemic philosophy of knowledge : from social facts to data, information and knowledge.* Kybernetes, 2002.

[Jurafsky & Martin 2013] Daniel Jurafsky et James H Martin. Speech and language processing : Pearson new international edition. Pearson, 2013.

[Jurgens *et al.* 2015] David Jurgens, Tyler Finethy, James McCorriston, Yi Tian Xu et Derek Ruths. *Geolocation prediction in twitter using social networks : A critical analysis and review of current practice.* In Ninth international AAAI conference on web and social media, 2015.

[Kannan *et al.* 2014] Subbu Kannan, Vairaprakash Gurusamy, S Vijayarani, J Ilamathi, M Nithya, S Kannan et V Gurusamy. *Preprocessing techniques for text mining.* International Journal of Computer Science & Communication Networks, vol. 5, no. 1, pages 7–16, 2014.

[Kanoje *et al.* 2015] Sumitkumar Kanoje, Sheetal Girase et Debajyoti Mukhopadhyay. *User profiling trends, techniques and applications.* arXiv preprint arXiv :1503.07474, 2015.

[Karsenty *et al.* 2001] L Karsenty, M et Zacklad et M Grundstein.

Capitaliser le contexte des décisions en conception : pourquoi et comment. Système d'Information pour la capitalisation des connaissances. Paris : Hermès, 2001.

[Kemighan *et al.* 1990] Mark D Kemighan, Kenneth Church et William A Gale. *A spelling correction program based on a noisy channel model.* In COLING 1990 Volume 2 : Papers presented to the 13th International Conference on Computational Linguistics, 1990.

[Kessler 2009] Rémy Kessler. *Traitement automatique d'informations appliqué aux ressources humaines.* Avignon, 2009.

[Kim *et al.* 2002] Sang-Bum Kim, Hae-Chang Rim, DongSuk Yook et Heui-Seok Lim. *Effective methods for improving naive bayes text classifiers.* In Pacific rim international conference on artificial intelligence, pages 414–423. Springer, 2002.

[Kim *et al.* 2017] Han Kyul Kim, Hyunjoong Kim et Sungzoon Cho. *Bag-of-concepts : Comprehending document representation through clustering words in distributed representation.* Neurocomputing, vol. 266, pages 336–352, 2017.

[Klement 2015] Joachim Klement. *Investor risk profiling : an overview.* 2015.

[Kocsis 2006] Richard N Kocsis. What is criminal profiling ? Springer, 2006.

[Kotsiantis *et al.* 2007] Sotiris B Kotsiantis, I Zaharakis, P Pintelas*et al. Supervised machine learning : A review of classification techniques.* Emerging artificial intelligence applications in computer engineering, vol. 160, no. 1, pages 3–24, 2007.

[Kuflik *et al.* 2003] Tsvi Kuflik, Bracha Shapira et Peretz Shoval. *Stereotype-based versus personal-based filtering rules in information filtering systems.* Journal of the American Society for Information Science and Technology, vol. 54, no. 3, pages 243–250, 2003.

[Kumar & Bhatia 2013] Lokesh Kumar et Parul Kalra Bhatia. *Text mining : concepts, process and applications.* Journal of Global Research in Computer Science, vol. 4, no. 3, pages 36–39, 2013.

[Lafferty *et al.* 2001] John Lafferty, Andrew McCallum et Fernando CN Pereira. *Conditional random fields : Probabilistic models for segmenting and labeling sequence data.* 2001.

[Laflaquière 2009] Julien Laflaquière. *Conception de système à base de traces numériques pour les environnements informatiques documentaires.* PhD thesis, Université de Technologie de Troyes, 2009.

[Lashkari *et al.* 2019] Arash Habibi Lashkari, Min Chen et Ali A Ghorbani. *A survey on user profiling model for anomaly detection in cyberspace.* Journal of Cyber Security and Mobility, pages 75–112, 2019.

[Lee 2001] Michael D Lee. *Fast text classification using sequential sampling processes.* In Australian Joint Conference on Artificial Intelligence, pages 309–320. Springer, 2001.

[Li *et al.* 2010] Feifei Li, Ke Yi et Wangchao Le. *Top-k queries on temporal data.* The VLDB journal, vol. 19, no. 5, pages 715–733, 2010.

[Liddy 2001] Elizabeth D Liddy. *Natural language processing.* 2001.

[Lund & Mille 2009] Kris Lund et Alain Mille. *Traces, traces d'interactions, traces d'apprentissages : définitions, modèles informatiques, structurations, traitements et usages.* Analyse de traces et personnalisation des environnements informatiques pour l'apprentissage humain. Hermès, pages 21–66, 2009.

[Mahesh 2020] Batta Mahesh. *Machine Learning Algorithms-A Review.* International Journal of Science and Research (IJSR).[Internet], vol. 9, pages 381–386, 2020.

[Malvache & Prieur 1993] Pierre Malvache et Pascal Prieur. *Mastering corporate experience with the Rex method.* In Proceedings of ISMICK, volume 93, pages 33–41, 1993.

[Martinet *et al.* 2002] Jean Martinet, Yves Chiaramella et Philippe Mulhem. *Un modèle vectoriel étendu de recherche d'information adapté aux images.* In INFORSID, volume 2, pages 337–348, 2002.

[Marx & Reichman 1984] Gary T Marx et Nancy Reichman. *Routinizing the discovery of secrets : Computers as informants.* American Behavioral Scientist, vol. 27, no. 4, pages 423–452, 1984.

[Matta *et al.* 2013a] Nada Matta, Guillaume Ducellier et Chaker Djaiz. *Traceability and structuring of cooperative Knowledge in design using PLM.* Knowledge Management Research & Practice, vol. 11, no. 1, pages 53–61, 2013.

[Matta *et al.* 2013b] Nada Matta, Guillaume Ducellier*et al. Memory Meetings-An Approach to Keep Track of Project Knowledge in Design.* In KDIR/KMIS, pages 336–343, 2013.

[Matta *et al.* 2016] Nada Matta, Hassan Atifi et Guillaume Ducellier. Daily knowledge valuation in organizations : Traceability and capitalization. John Wiley & Sons, 2016.

[Merialdo 1995] Bernard Merialdo. *Modèles probabilistes et étiquetage automatique.* TAL. Traitement automatique des langues, vol. 36, no. 1-2, pages 7–22, 1995.

[Merzeau 2013] Louise Merzeau. *L'intelligence des traces.* Intellectica-La revue de l'Association pour la Recherche sur les sciences de la Cognition (ARCo), vol. 1, no. 59, pages p–115, 2013.

[Mille 2013] Alain Mille. *De la trace à la connaissance à l'ère du Web. Introduction au dossier.* Intellectica, vol. 59, no. 1, pages 7–28, 2013.

[Miloslavskaya & Tolstoy 2016] Natalia Miloslavskaya et Alexander Tolstoy. *Big data, fast data and data lake concepts.* Procedia Computer Science, vol. 88, pages 300–305, 2016.

[Moens 2000] Marie-Francine Moens. *Automatic indexing and abstracting of document texts, Series : The Information Retrieval Series, vol. 6*, 2000.

[Mohamadally & Fomani 2006] H Mohamadally et B Fomani. *SVM : Machines a vecteurs de support ou separateurs a vastes marges.* Survey, Versailles St Quentin, vol. 16, 2006.

[Nahm & Mooney 2002] Un Yong Nahm et Raymond J Mooney. *Text mining with information extraction.* In Proceedings of the AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, pages 60–67. Stanford CA, 2002.

[NINTEX] NINTEX. definitive guide to America's Most Broken Processes, *howpublished = https://www.nintex.com, note = Accessed : 2020-06-30.*

[Nonaka & Konno 1998] Ikujiro Nonaka et Noboru Konno. *The concept of "Ba" : Building a foundation for knowledge creation.* California management review, vol. 40, no. 3, pages 40–54, 1998.

[Nonaka *et al.* 1995] Ikujirō Nonaka, Ikujir o Nonaka, Nonaka Ikujiro, Hirotaka Takeuchi*et al.* The knowledge-creating company : How japanese companies create the dynamics of innovation, volume 105. OUP USA, 1995.

[Orkin & Roy 2010] Jeff Orkin et Deb Roy. *Semi-automated dialogue act classification for situated social agents in games.* In International Workshop on Agents for Games and Simulations, pages 148–162. Springer, 2010.

[Ortigosa *et al.* 2014] Alvaro Ortigosa, Rosa M Carro et José Ignacio Quiroga. *Predicting user personality by mining social interactions in Facebook.* Journal of computer and System Sciences, vol. 80, no. 1, pages 57–71, 2014.

[Osisanwo *et al.* 2017] FY Osisanwo, JET Akinsola, O Awodele, JO Hinmikaiye, O Olakanmi et J Akinjobi. *Supervised machine learning algorithms : classification and comparison.* International Journal of Computer Trends and Technology (IJCTT), vol. 48, no. 3, pages 128–138, 2017.

[Polanyi 1961] Michael Polanyi. *Knowing and being.* Mind, pages 458–470, 1961.

[Poo *et al.* 2003] Danny Poo, Brian Chng et Jie-Mein Goh. *A hybrid approach for user profiling.* In 36th Annual Hawaii International Conference on System Sciences, 2003. Proceedings of the, pages 9–pp. IEEE, 2003.

[Popa *et al.* 2011] Iulian Sandu Popa, Karine Zeitouni, Vincent Oria, Dominique Barth et Sandrine Vial. *Indexing in-network trajectory flows.* The VLDB Journal, vol. 20, no. 5,

pages 643–669, 2011.

[Prax 2012] Jean-Yves Prax. Manuel du knowledge management-3ème édition : Mettre en réseau les hommes et les savoirs pour créer de la valeur. Hachette, 2012.

[Ramesh & Edwards 1993] Balasubramaniam Ramesh et Michael Edwards. *Issues in the development of a requirements traceability model.* In [1993] Proceedings of the IEEE International Symposium on Requirements Engineering, pages 256–259. IEEE, 1993.

[Rauscher 2016] François Rauscher. *Gestion des connaissances et communication médiatisée : traçabilité et structuration des messages professionnels.* PhD thesis, Troyes, 2016.

[Richard 1992] Jean-François Richard. *Les activités mentales. Comprendre, raisonner, trouver des solutions.* Revue Philosophique de la France Et de l, vol. 182, no. 4, 1992.

[Rish *et al.* 2001] Irina Rish, Joseph Hellerstein et Jayram Thathachar. *An analysis of data characteristics that affect naive Bayes performance.* IBM TJ Watson Research Center, vol. 30, pages 1–8, 2001.

[Roche 2005] Christophe Roche. *Terminologie et ontologie.* Langages, no. 1, pages 48–62, 2005.

[Rodríguez-García *et al.* 2014] Miguel Ángel Rodríguez-García, Rafael Valencia-García, Francisco García-Sánchez et J Javier Samper-Zapater. *Creating a semantically-enhanced cloud services environment through ontology evolution.* Future Generation Computer Systems, vol. 32, pages 295–306, 2014.

[Romo & Capiluppi 2015] Bilyaminu Auwal Romo et Andrea Capiluppi. *Towards an automation of the traceability of bugs*

from development logs : a study based on open source software. In Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering, pages 1–6, 2015.

[Roul 2018] Rajendra Kumar Roul. *An effective approach for semantic-based clustering and topic-based ranking of web documents.* International Journal of Data Science and Analytics, vol. 5, no. 4, pages 269–284, 2018.

[Sakr & Al-Naymat 2010] Sherif Sakr et Ghazi Al-Naymat. *Graph indexing and querying : a review.* International Journal of Web Information Systems, 2010.

[Salperwyck & Lemaire 2011] Christophe Salperwyck et Vincent Lemaire. *Learning with few examples : An empirical study on leading classifiers.* In The 2011 international joint conference on neural networks, pages 1010–1019. IEEE, 2011.

[Sauban & Pfahringer 2003] Maximilien Sauban et Bernhard Pfahringer. *Text Categorisation Using Document Profiling.* In Knowledge Discovery in Databases : PKDD 2003 : 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003, Proceedings, volume 7, page 411. Springer Science & Business Media, 2003.

[Schiaffino & Amandi 2000] Silvia N Schiaffino et Analia Amandi. *User profiling with Case-Based Reasoning and Bayesian Networks.* In IBERAMIA-SBIA 2000 open discussion track, pages 12–21. Citeseer, 2000.

[Schneider 2005] Karl-Michael Schneider. *Techniques for improving the performance of naive bayes for text classification.* In International Conference on Intelligent Text Processing

and Computational Linguistics, pages 682–693. Springer, 2005.

[Schreiber *et al.* 2018] Marc Schreiber, Bodo Kraft et Albert Zündorf. *NLP Lean Programming Framework : Developing NLP Applications More Effectively.* In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Demonstrations, pages 1–5, 2018.

[Segonds 2011] Frédéric Segonds. *Contribution to the integration of a collaborative design environment in the early stages of design.* PhD, Arts et Metiers ParisTech, 2011.

[Sellah 2019] Smail Sellah. *Approche automatisée d'assistance à la structuration des connaissances.* PhD thesis, Bourgogne Franche-Comté, 2019.

[Serres 2002] Alexandre Serres. *Quelle (s) problématique (s) de la trace ?* 2002.

[Shardanand & Maes 1995] Upendra Shardanand et Pattie Maes. *Social information filtering : Algorithms for automating "word of mouth".* In Proceedings of the SIGCHI conference on Human factors in computing systems, pages 210–217, 1995.

[Sheth 1994] Beerud Dilip Sheth. *A learning approach to personalized information filtering.* PhD thesis, Massachusetts Institute of Technology, 1994.

[Shieh & Keogh 2008] Jin Shieh et Eamonn Keogh. *i SAX : indexing and mining terabyte sized time series.* In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 623–631, 2008.

[Silberztein 1993] Max Silberztein. Dictionnaires électroniques et analyse automatique de textes : le système intex. Masson, 1993.

[Silberztein 2016] Max Silberztein. Formalizing natural languages : The nooj approach. John Wiley & Sons, 2016.

[Smyth & Cotter 2001] Barry Smyth et Paul Cotter. *Personalized electronic program guides for digital TV.* Ai Magazine, vol. 22, no. 2, pages 89–89, 2001.

[Solangi *et al.* 2018] Yasir Ali Solangi, Zulfiqar Ali Solangi, Samreen Aarain, Amna Abro, Ghulam Ali Mallah et Asadullah Shah. *Review on Natural Language Processing (NLP) and its toolkits for opinion mining and sentiment analysis.* In 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS), pages 1–4. IEEE, 2018.

[Soltysiak & Crabtree 1998] SJ Soltysiak et IB Crabtree. *Automatic learning of user profiles—towards the personalisation of agent services.* BT Technology Journal, vol. 16, no. 3, pages 110–117, 1998.

[Stefanidis *et al.* 2018] Kostas Stefanidis, Eirini Ntoutsi, Haridimos Kondylakis et Yannis Velegrakis. *Social-Based Collaborative Filtering.*, 2018.

[Straka & Straková 2017] Milan Straka et Jana Straková. *Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe.* In Proceedings of the CoNLL 2017 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies, pages 88–99, 2017.

[Subrahmanian *et al.* 2005] Eswaran Subrahmanian, Sudarsan Rachuri, Steven J Fenves, Sebti Foufou et Ram D Sriram. *Product lifecycle management support : a challenge in*

supporting product design and manufacturing in a networked economy. International Journal of Product Lifecycle Management, vol. 1, no. 1, pages 4–25, 2005.

[Sun *et al.* 2017] Shiliang Sun, Chen Luo et Junyu Chen. *A review of natural language processing techniques for opinion mining systems.* Information fusion, vol. 36, pages 10–25, 2017.

[Tablan *et al.* 2004] Valentin Tablan, Diana Maynard, Kalina Bontcheva, Hamish Cunningham, V Tablan, D Maynard et K Bontcheva. *Gate, an Application Developer's Guide.* Department of Computer Science, University of Sheffield, UK, vol. 19, 2004.

[Tiercelin 2012] Claudine Tiercelin. *Métaphysique et philosophie de la connaissance.* L'annuaire du Collège de France. Cours et travaux, no. 111, pages 657–679, 2012.

[Tourtier 1995] Paul-André Tourtier. *Analyse préliminaire des métiers et de leurs interactions.* Rapport intermédiaire du projet GENIE, INRIA-Dassault-Aviation, 1995.

[Tsoumakas *et al.* 2009] Grigorios Tsoumakas, Ioannis Katakis et Ioannis Vlahavas. *Mining multi-label data.* In Data mining and knowledge discovery handbook, pages 667–685. Springer, 2009.

[Tsuchiya *et al.* 2013] Ryosuke Tsuchiya, Tadahisa Kato, Hironori Washizaki, Masumi Kawakami, Yoshiaki Fukazawa et Kentaro Yoshimura. *Recovering traceability links between requirements and source code in the same series of software products.* In Proceedings of the 17th International Software Product Line Conference, pages 121–130, 2013.

[Tsuchiya *et al.* 2015] Ryosuke Tsuchiya, Hironori Washizaki, Yoshiaki Fukazawa, Keishi Oshima et Ryota Mibe. *Interactive*

recovery of requirements traceability links using user feedback and configuration management logs. In International Conference on Advanced Information Systems Engineering, pages 247–262. Springer, 2015.

[Turban 2013] Bernhard Turban. Tool-based requirement traceability between requirement and design artifacts. Springer Science & Business Media, 2013.

[Van Otterlo 2013] Martijn Van Otterlo. *A machine learning view on profiling.* Privacy, Due Process and the Computational Turn-Philosophers of Law Meet Philosophers of Technology. Abingdon : Routledge, pages 41–64, 2013.

[Vapnik 1999] Vladimir N Vapnik. *An overview of statistical learning theory.* IEEE transactions on neural networks, vol. 10, no. 5, pages 988–999, 1999.

[Vicient *et al.* 2011] Carlos Vicient, Antonio Moreno*et al.* *Ontology-based feature extraction.* In 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, volume 3, pages 189–192. IEEE, 2011.

[Wang & Koopman 2017] Shenghui Wang et Rob Koopman. *Clustering articles based on semantic similarity.* Scientometrics, vol. 111, no. 2, pages 1017–1031, 2017.

[Webb *et al.* 2010] Geoffrey I Webb, Eamonn Keogh et Risto Miikkulainen. *Naïve Bayes.* Encyclopedia of machine learning, vol. 15, pages 713–714, 2010.

[Wong & Raghavan 1984] SK Michael Wong et Vijay V Raghavan. *Vector space model of information retrieval : a reevaluation.* In SIGIR, volume 84, pages 167–185, 1984.

[Yang 2010] Yinghui Catherine Yang. *Web user behavioral profiling for user identification.* Decision Support Systems, vol. 49, no. 3, pages 261–271, 2010.

[Ying *et al.* 2018] Qiu Fang Ying, Dah Ming Chiu, Srinivasan Venkatramanan et Xiaopeng Zhang. *User modeling and usage profiling based on temporal posting behavior in OSNs.* Online Social Networks and Media, vol. 8, pages 32–41, 2018.

[Yvon 2010] François Yvon. *Une petite introduction au traitement automatique des langues naturelles.* In Conference on Knowledge discovery and data mining, pages 27–36, 2010.

Elamin ABDERRAHIM

Doctorat : Systèmes SocioTechniques

Année 2022

Know-linking : favoriser un partage systématique et ciblé des connaissances entre les acteurs de l'entreprise

La connaissance est un capital qui a de la valeur pour l'entreprise, ce capital peut être perdu suite à plusieurs raisons, tels que le départ des experts qui ne laissent pas des traces, ou la perte d'accès à certaines ressources. Le besoin de réutiliser l'expérience passée et de partager des connaissances pour la réalisation des projets est devenu une exigence. L'objectif des entreprises depuis plusieurs années, est alors de construire une stratégie réussie pour gérer et partager les connaissances, surtout dans un contexte qui a subi une évolution importante. Les méthodes traditionnelles ont montré leurs limites à l'ère de l'explosion de données et de l'évolution technologique. Dans ce contexte, nous proposons notre projet de thèse, fruit d'une étude bibliographique qui a conduit à l'élaboration de l'approche *Know-linking*. Une approche composée de trois étapes, dont la première consiste à profiler les collaborateurs de l'entreprise afin d'analyser leurs besoins en connaissances. Dans la deuxième étape, nous générons une représentation des profils sous forme des graphes et nous fouillons les documents pour retrouver les liens sémantiques entre ces profils. La troisième étape est une étape de distribution de supports de connaissances selon les profils et liens entre eux. Notre approche favorise une génération de connaissances partagées d'une façon personnalisée (par profil de collaborateur) et systématique (assurée par un système). Nos travaux ont conduit à la construction d'une infrastructure de test de l'approche.

Mots clés : échange de connaissances – traçabilité – profilage (droit) – traitement automatique du langage naturel – gestion des documents.

Know-linking: Promote Systematic and Personalised Knowledge Sharing between Actors in a Company

Knowledge is a valuable asset for the company, which can be lost for many reasons, such as the departure of experts who do not leave a trace, or the loss of access to certain resources. The need to reuse past experience and share knowledge for project implementation has become a requirement. The goal of companies for several years has been to build a successful strategy to manage and share knowledge, especially in a context that has undergone significant change. Traditional methods have shown their limits in the era of data explosion and technological evolution. In this context, we propose this thesis project, which is the result of a bibliographical study that led to the development of the Know-linking approach. This approach consists of three steps, the first of which consists of profiling the company's employees in order to analyze their knowledge needs. In the second step, we generate a representation of the profiles in the form of graphs and we search the documents to find the semantic links between these profiles. The third step is a distribution of knowledge supports according to the profiles and links between them. Our approach favors the generation of shared knowledge in a personalized (by employee profile) and systematic (provided by a system) way. Our work led to the development of a framework for the approach.

Keywords: knowledge sharing – traceability – natural language processing (computer science) – records management.

Thèse réalisée en partenariat entre :

ANDRA

utt UNIVERSITÉ DE TECHNOLOGIE TROYES

Ecole Doctorale "Sciences pour l'Ingénieur"