Smart Zeenea : Services Intelligents pour la Maintenance et l'Enrichissement de Catalogues de Données

Projet Doctoral

Laboratoire : LIP6 – Sorbonne Université

Entreprise : Zeenea

Septembre 2023

1 Contexte et problématique générale

1.1 Gestion d'informations dans l'ère du Big Data

Les entreprises sont de plus en plus confrontées à des besoins d'organisation, de gestion et de partage d'informations conservées sous divers formats par des systèmes et sur des supports autonomes et divers. L'hétérogénéité et l'évolutivité des formats, systèmes et supports nécessitent de nouvelles solutions pour faciliter l'utilisation et le partage pérenne de ces données.

Dans ce contexte, nous pouvons distinguer entre trois types de systèmes d'intégration de données qui répondent à des besoins complémentaires :

- Un *entrepôt de données* (data warehouse) [3] est un système qui sert à collecter des données brutes provenant de différents datasets (bases de données, fichiers, systèmes d'informations, ...), de les transformer en données structurées et de les entreposer ensuite dans une base de données multi-dimensionnelle. Les données obtenues par ce processus Extract-Transform-Load (ETL) peuvent ensuite être organisées et analysées grâce à des requêtes analytiques (OLAP) et des outils de visualisation. L'organisation des données dans un

entrepôt de données nécessite la définition de *schémas analytiques* qui sont orientés vers des domaines et qui répondent à des besoins d'analyse bien définis.

- Un *système polystore* [2] répond aux besoins d'applications qui doivent interagir simultanément avec plusieurs "îles d'informations" réunissant des systèmes de données avec un modèle partagé (relationnel, noSQL, flux/stream). Les services polystore aident à fédérer ces îles d'informations grâce à des requêtes "espéranto" et des architectures "middleware" qui traduisent, optimisent et exécutent les requêtes fédérées.

- L'objectif d'un *datalake* [6] est de rassembler tout le *patrimoine de données* d'une entreprise ou d'une organisation sous forme d'une collection de *datasets* autonomes qui peuvent représenter des bases de données structurées, des collections de documents XML, des données JSON et d'autres fichiers texte, images ou audio. Ces datasets peuvent être accompagnées de métadonnées (schémas, statistiques, descriptions textuelles, fingerprint). Le rôle d'un *catalogue de données* est d'organiser cet espace afin de faciliter la recherche et l'intégration de datasets par rapport à des besoins spécifiques.

Les trois types de systèmes/architectures proposent des services complémentaires de création, d'intégration et d'exploitation d'informations (Tableau 1). La complexité des services fournis par chaque système est généralement proportionnelle à la complexité et à l'effort nécessaire pour l'acquisition et l'intégration d'un nouveau dataset. Un système entrepôt de données propose des services avancés d'*analyse de données* structurées (requêtes analytiques) et nécessite un effort important de modélisation et d'intégration de datasets (chaque nouveau dataset nécessite la définition d'un ou de plusieurs processus ETL). Un système polystore facilite la *création de nouveaux datasets* à partir des datasets stockées dans des systèmes hétérogènes (requêtes fédérées) grâce à une architecture middleware et des procédures de transformation/réplication des données entre différents systèmes. Enfin, un système datalake permet d'*organiser des grandes collections de datasets autonomes (datalake) et d'identifier des datasets* pour des besoins d'information spécifiques grâce à un catalogue de données (requêtes catalogues de métadonnées).

	Datalake	**Polystore**	**Entrepôt**
Modèle	catalogues de métadonnées	schémas et requêtes fédérés	schémas et requêtes analytiques
Services	organisation et découverte de datasets	requêtes fédérées sur des bases de données autonomes	analyse et visualisation de données multidimensionnelles
Tâches / Effort	extraction et indexation de métadonnées	intégration de systèmes autonomes	extraction, transformation et stockage de données

TAB. 1 : Datalakes, polystores et entrepôts de données

1.2 La solution Zeenea

La société Zeenea fournit à ces clients des outils pour l'exploration et l'exploitation de leur patrimoine de données en utilisant une solution basée entièrement sur le *cloud* (SAAS) :

Le service *Zeenea Studio* permet de gérer, maintenir et enrichir la documentation du patrimoine de données d'un client sous forme d'un catalogue de données (Section 1.3). Le catalogue est un inventaire de divers types d'objets comme les datasets, les processus de données, la terminologie métier (business terms), les visualisations, etc. Il facilite l'enrichissement de la documentation et la détection de liens (similarité, inclusion) entre des d'objets pour améliorer la contextualisation sémantique des datasets et faciliter leur utilisation à travers un graphe de connaissance.

Le service *Zeenea Explorer* permet la construction et la maintenance du data catalogue. L'extraction des métadonnées est effectuée par des *scanners* déployés. Les scanners permettent de charger des métadonnées issues des différentes bases de données du client telles que les schémas de données (noms et descriptions textuelles des tables et colonnes), des statistiques sur les données (taille, distribution), des documentions, etc. Zeenea Explorer trouve via son moteur de recherche des datasets pertinents selon le profil et les préférences des utilisateurs et selon les différents concepts du catalogue de données (Section 1.3). Le service effectue aussi une priorisation des informations en fonction des attentes et besoins des différentes équipes et type d'utilisateur (un data scientist et un chargé de marketing n'ont pas le même intérêt sur les données). Enfin, Zeenea Explorer facilite le partage des connaissances sur les différents items du catalogue de données et de leurs cas d'usages à l'aide de fonctionnalités collaboratives diverses et un accès à une documentation qui est mise à jour en temps réel.

1.3 Le catalogue de données Zeenea

Les catalogues de données sont au cœur du fonctionnement des services Zeenea pour représenter les connaissances spécifiques au métier du client et les liens avec les datasets à organiser. La Figure 1 représente l'architecture détaillée du catalogue de donnée Zeenea :

- Le *Schéma Métier* (business schema) est un graphe de *concepts métier* (business concepts) pour la description structurée des *objets métier* (business items). En pratique, un schéma métier est souvent une hiérarchie de concepts, mais il est possible de créer des structures plus complexes (graphes de concepts).
- Le *Glossaire Métier* (business glossary) définit les *termes métiers* (business terms) utilisés comme valeurs des propriétés de concepts. Les glossaires proposent différentes vues (fonctionnelle, technique, administratives, légales) sur les informations de l'entreprise.
- *Objets métier* (business item) : Le métaschéma et le glossaires sont utilisés pour définir des objets métiers (par exemple, un composant technique d'une voiture, un type de client, etc).
- La notion de *Dataset* est abstraite et peut identifier une ou plusieurs colonnes dans une table, une ou plusieurs tables, un répertoire avec des images, etc. Dans un catalogue de données, un dataset est représenté par un ensemble de métadonnées générées par les

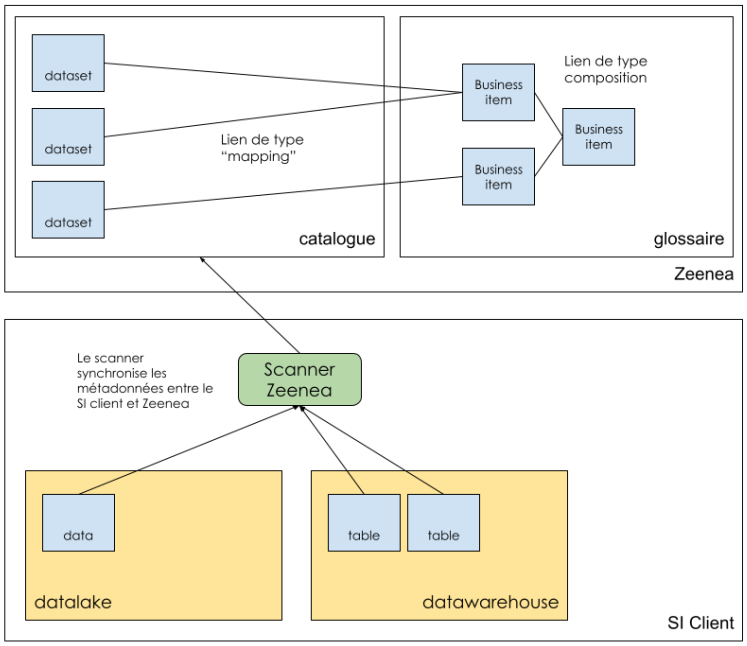

FIG. 1 : Modèle du catalogue de données

scanners Zeenea. Ces métadonnées permettent d'identifier les ressources dans le catalogue et décrivent le contenu de ces ressources (descriptions textuelles, statistiques, etc).

- *Mappings* (linked items) : Les mappings relient les datasets (métadonnées) aux objets métiers d'un catalogue de données.

2 Enjeux pour la société Zeenea

La transformation digitale des organisations a eu pour effet de démultiplier le nombre de jeux de données produits, ainsi que le nombre de personnes impliquées dans leur consommation et leur production. Mais les organisations manquent de moyens pour identifier, localiser et comprendre ces jeux de données, qui restent par conséquents opaques pour le plus grand nombre, et leur potentiel ne peut être exploité. Afin d'illustrer les défis rencontrés dans la construction d'un catalogue Zeenea, nous pouvons donner l'exemple d'un catalogue qui est en construction pour un grand établissement financier. Cet établissement a commencé à cataloguer huit de ces systèmes de stockage de données qui comprend un datalake avec environ 240 000 datasets. Actuellement 3252 datasets et les 54 663 champs qui les composent ont été documentés dans le catalogue Zeenea. Pour être considéré comme documenté, chacun de ces objets du catalogue a, en plus d'un nom compréhensible et une description obligatoire, une quinzaine de propriétés documentées. Ces propriétés, différentes pour chaque catalogue, représentent la plupart du temps diverses classifications dépendantes de l'organisation telles que la confidentialité, la sensibilité, les entités propriétaires de la donnée, etc. Pour compléter la documentation, les différents objets du catalogue représentant les données doivent également être associés aux objets métiers (programmes, processus) qui les produisent et qui les utilisent. En tout, l'exemple présenté, représente près d'un million d'objets à documenter, ce qui nécessite des nouvelles solutions pour *l'automatisation* de certaines tâches dans la construction du catalogue. Dans cette perspective,

la société Zeenea est actuellement en train de définir la prochaine version du système que nous allons appeler *Smart Zeenea*.

2.1 Smart Zeenea

Une gestion déclarative et manuelle des catalogues de données ne permet pas d'intégrer un nombre croissant de jeux de données pour répondre aux besoins d'un nombre grandissant d'utilisateurs et avec un cycle de vie de plus en plus dynamique. Dans ce contexte, l'utilisation de l'intelligence artificielle dans le catalogue de données permettra d'automatiser certaines tâches pour réduire les couts de maintenance d'un catalogue et de démocratiser l'usage de la donnée.

Pour répondre à ces besoins d'automatisation, la société Zeenea a pour objectif d'étendre le produit actuel avec de nouveaux services intelligents :

1. **Détection de données sensibles et de données personnelles :** Avec une pression accrue sur la conformité aux diverses réglementations relatives à la confidentialité et à la sécurité des données [7], les Data Stewards recherchent des outils qui peuvent automatiser les informations personnelles identifiables (PII). Les entreprises ont besoin également d'identifier facilement leurs données sensibles afin de maîtriser leur diffusion et leur utilisation.

2. **Versionnement, déduplication, alignement d'informations :** Les organisations stockent souvent la même information, dans des systèmes divers et sous différents formats. La détection d'informations dupliquées ou similaires est un moyen important pour accélérer la documentation pour les Data Stewards, améliorer la recommandation de datasets pour les utilisateurs finaux, et optimiser le cycle de vie et le stockage des données.

3. **Organisation par thématiques :** Différents datasets peuvent contenir des informations concernant le même domaine ou la même thématique du catalogue de données. L'identification automatique de ces thématiques réduit l'effort de maintenance du catalogue et améliore la recherche "sémantique" de datasets pour les utilisateurs.

4. **Suivi d'évolution des datasets :** De nombreux datasets sont obtenus par des processus de transformation de données. Pouvoir retracer l'évolution de données jusqu'à leurs origines et examiner l'évolution de leur cycle de vie facilite l'amélioration de la qualité des données et des processus dans une entreprise et augmente la confiance et la transparence des processus de décision.

5. **Extraction de résumés, motifs, statistique** : La reconnaissance automatique des structures, des propriétés statistiques et d'autres motifs réguliers présents dans les datasets permet d'enrichir le catalogue de données pour donner aux Data Stewards et aux utilisateurs une vision plus globale de l'information.

6. **Moteur de recherche sémantique et personnalisé :** Enfin, la quantité et la richesse des informations représentées dans le catalogue de données ouvre la possibilité de construire un moteur de recherche pour retrouver des objets pertinents, en prenant en compte le contexte et l'historique des activités d'un utilisateur ou d'un groupe d'utilisateurs.

3 Projet doctorale

3.1 Objectifs

L'objectif de ce projet doctoral est d'étendre le système Zeenea avec de nouvelles fonctionnalités pour assister les concepteurs d'un catalogue Zeenea. La motivation est de construire des catalogues intelligents qui sont capables de s'adapter automatiquement à l'évolution d'un datalake et de faciliter l'exploration de ses datasets par de nouveaux services intelligents pour la génération et la maintenance de documentation, la classification sémantique des datasets (par rapport au glossaire métier), la classification et détection de données personnelles (Personally Identifiable Information PII) et la prédiction et suggestion de liens (linked-items) entre des datasets et les différents objets métiers.

On peut distinguer plusieurs tâches fondamentales dans la construction et l'enrichissement du catalogue de données Zeenea :

- *Enrichissement du glossaire métier :* La difficulté principale de cette tâche est la définition du glossaire qui nécessite une bonne connaissance du domaine applicatif à modéliser. Des entreprises spécialisées vendent des glossaires pour différents domaines (automobile, assurance, etc). Ces glossaires sont souvent organisés sous forme hiérarchique et doivent être adaptés au schéma du catalogue et au contexte spécifique des utilisateurs. Pour avoir un glossaire métier adapté aux données, une première approche consiste à découvrir des concepts métiers à partir des métadonnées. Une autre solution consiste à utiliser les concepts déjà présents dans le glossaire métiers pour en découvrir des nouveaux grâce à des techniques de génération et enrichissement de textes.
- *Suggestion de linked-items* : Le rôle principal du catalogue Zeenea est de relier les métadonnées sur les datasets aux objets métiers du catalogue grâce à des mappings (linked-items). On suppose que le glossaire existe et l'objectif est d'assister les experts dans la génération des alignements (linked-items) entre les datasets et les concepts. On a déjà identifié plusieurs types de relations entre les datasets et les concepts métier qui peuvent être utilisés pour la découverte de nouveaux linked-items. Une première implémentation qui exploite la similarité linguistique et sémantique entre les descriptions des datasets et des objets métiers a été effectuée avec succès dans le cadre du stage Master du candidat Anis Aknouche [1]. D'autres relations à explorer sont (1) l'appartenance des datasets au *domaine* d'un concept métier, (2) les relations d'inclusion de datasets dans d'autres datasets déjà alignés, (3) le chevauchement des domaines d'attributs de différents datasets, (4) les liens de provenance et dépendance entre les datasets, ainsi que (5) les liens vers des concepts ou attributs représentant des données personnelles (adresse, email) pour la détection des données personnelles (PII).
- *Enrichissement des propriétés des objets du catalogue de données* : Les objets métiers sont définis par les experts pour enrichir le schéma du catalogue et le glossaire métiers. Tous les objets métiers contiennent une description textuelle et d'autres propriétés définies par les experts du domaine. Une première approche est l'enrichissement des descriptions textuelles des objets métiers avec des topics extraits dans les descriptions (métadonnées) des datasets liés au concept métier [1]. Les métadonées des datasets peuvent aussi être enrichies en générant d'autres propriétés telles que des vecteurs "embeddings" pour les données textuelles et des distributions de valeurs pour les données numériques.

3.2 Approche et enjeux scientifiques

Il existe de nombreuses approches et solutions pour la création et l'enrichissement de catalogues de données. Dans le cadre du projet doctoral Smart Zeenea nous allons nous concentrer sur les enjeux suivants :

- **Modélisation du catalogue :** Le premier enjeu est de définir une représentation générique du catalogue pour pouvoir exploiter des techniques récentes en intelligence artificielle et étendre le système avec de nouveaux services *intelligents* (détection de PII, détection de similarité, etc). Modéliser le catalogue Zeenea sous forme de graphe permet de structurer et d'organiser les objets du catalogue de données de façon simple et pratique à exploiter [5]. Les concepts métiers et les datasets seront représentés par des nœuds avec différentes propriétés et reliés par les linked-items. Cette représentation sous forme de graphe facilite l'analyse et l'intégration des différents concepts et relations du catalogue de données et de mieux exploiter les algorithmes existants (détection de communités, classification et prédiction de nœuds et de liens) pour la classification, la recommandation et la découverte de concepts et de linked items.

- **Extraction et enrichissement de métadonnées** : En partant du modèle de catalogue, l'enjeu est d'automatiser le processus d'extraction des métadonnées. Nous envisageons d'explorer différentes pistes de solution utilisant, par exemple, des méthodes d'auto-tagging simples (regex) et des méthodes plus récentes fondées sur l'apprentissage non-supervisées (clustering) et supervisées (deep learning). La société Zeenea a déjà collaboré avec le laboratoire LIP6 dans le cadre de deux stages M2 sur le problème d'enrichissement automatique du catalogue avec des métadonnées textuelles et numériques :

  - **Métadonnées textuelles** ([1], Section 5) : Le modèle Zeenea est fondé sur les glossaires pour la description sémantique du contenu et des relations entre les datasets d'un catalogue de données. Les concepts sont définis par des descriptions textuelles pour faciliter la visualisation et l'exploration du data catalogue. Le défi est d'extraire, à partir des descriptions textuelles, des informations plus structurées pour pouvoir les exploiter dans la découverte d'alignements avec les métadonnées. Un stage sur ce sujet effectué par le candidat a déjà produit des premiers résultats prometteurs.
  - **Métadonnées numériques** ([4], Section 5) : Les métadonnées des objets du data catalogue peuvent aussi être de type numérique. Des méthodes statistiques et probabilistes peuvent être utilisées pour décrire le contenu des datasets et détecter des motifs en modélisant la distribution de valeurs pour des colonnes dans une table donnée. Une approche basée sur le fingerprint de données numériques des colonnes des datasets structurés a été effectuée dans le cadre du stage Master de Li Haoran [4].

- **Optimisation de ressources :** Notre objectif est d'exploiter des méthodes d'enrichissements de données récents et tirer profit des modèles de ML et DeepL pour la classification ou l'inférence de relations. Un enjeu particulier liée à l'architecture distribuée de Zeenea, est que les métadonnées sont produites avec les ressources de calcul des clients. Ceci limite l'application de certaines techniques d'extraction de métadonnées qui nécessitent, par exemple, l'apprentissage de modèles statistiques. Une solution est, par exemple, d'adapter des modèles pré-entrainés (Google word2vec ou Bert, et Stanford Glove pour les modèles de langue, ou Google Inception et Micorsoft ResNet pour la classification d'images) par

des techniques d'apprentissage par transfert (transfer learning) et d'enrichir les métadonnées obtenues en utilisant les informations contextuelles (alignements, objets métier) du catalogue.

4 Méthodologies de recherche

Nos travaux de recherche sont dirigés par les besoins industriels d'automatiser la construction de catalogues de données et nous allons appliquer et combiner plusieurs méthodologies de recherche :

1. Exploration et définition de cas d'usage : dans un premier temps, nous allons explorer plusieurs cas d'usage avec des données réelles pour préciser les hypothèses et choisir un cadre pratique (données, fonctionnalités) qui sera utilisé pour valider les solutions.
2. État de l'art scientifique et technique : en même temps, nous allons approfondir nos connaissances sur les méthodes scientifiques et sur les outils techniques (librairies, modèles) existantes pour résoudre les problèmes posés.
3. Formalisation et modélisation : nous allons étudier et comparer différentes approches de modélisation sémantiques de métadonnées sous forme de graphes de connaissances (RDF/OWL, graphes de propriétés) pour la représentation du data catalogue. En partant du graphe de catalogue, nous allons étudier et proposer des solutions pour l'enrichissement du catalogue (graphe de connaissances). Un aspect important réside dans l'élaboration des solutions fédérées pour optimiser l'utilisation des ressources de stockage et de calcul distribuées (client, cloud).
4. Prototypage, expérimentation et validation : Le prototypage et l'expérimentation représente une phase importante du travail à effectuer qui inclut le choix de données réelles (datasets, catalogues), le développement des solutions en utilisant des outils (librairies, modèles) existants, la définition de mesures pour évaluer la qualité des résultats et l'expérimentation finale.

5 Travaux antérieurs

Stage 1 : Approches NLP pour l'enrichissement de métadonnées dans un data-lake [1] Dans ce stage, deux approches ont été présentées pour la suggestion de linked-items entre les business-items et les colonnes d'un patrimoine de données. La première approche se base sur le clustering des vecteurs embeddings générés par la concaténation des noms de colonnes avec leurs descriptions textuelles dans le cas où elles sont disponibles. Le clustering est fait avec HDBSCAN précédé par une réduction de dimensions avec UMAP. La deuxième approche consiste à appliquer dans un premier temps une réduction de dimension avec UMAP suivie d'un clustering avec HDBSCAN sur les colonnes avec descriptions textuelles.

Les approches proposées peuvent être généralisées et appliquées pour générer des suggestions de linked-items entre les autres concepts métiers du catalogue de données et les tables et datasets du patrimoine de données. Cela permettra aussi de générer et d'enrichir les descriptions

textuelles des différents objets du catalogue de données grâce aux descriptions textuelles des topics auxquels ils appartiennent ou qui leur sont similaires sémantiquement.

Stage 2 : Approche FINGERPRINT pour l'enrichissement de métadonnées dans un datalake [4] L'utilisation de méthodes NLP pour aligner des colonnes d'une table est limitée par l'existence de noms de colonnes significatifs (par exemple, adresse, ville) et de descriptions textuelles. Le but de ce stage était de déterminer les types sémantiques de colonnes à partir de leur contenu. Nous avons étudié différentes solutions de génération de *fingerprints* pour comparer et aligner des colonnes numériques (percentiles) et textuelles (embeddings, ensemblistes) dans des données tabulaires. Cette première étude nous a permis de mieux comprendre les défis et nous envisageons d'étudier d'autres pistes d'annotation sémantique dans le cadre de ce projet doctoral.

6 Moyens et Matériel

Pour assurer le bon déroulement de la thèse, l'entreprise Zeenea et le LIP6 s'engagent à fournir les moyens et le matériel nécessaires au candidat. Coté Zeenea, le candidat sera intégrer à une équipe R&D responsable du développement du produit, et aura en sa possession un MacBook Professionnel, un accès à une version de développement du Data Catalog Zeenea ainsi qu'à des instances de calcul et de stockage sur le cloud (AWS, GCP, etc) ou toute autre technologie jugée nécessaire pour l'avancement du projet. Coté LIP6, le candidat sera amené à intégrer l'équipe Base de Données (BD) et aura en sa possession un ordinateur de bureau et un accès aux serveurs de la PPTI (Plateforme Pédagogique et Technique Informatique).

Dans le cadre de ce projet de thèse, la validation expérimentale des solutions développées sur des données réelles est d'une grande importance. Pour valider les méthodes et les approches proposées, nous devrons étudier différents scénarios pour mettre en place un environnement expérimental avec un catalogue test. Dans un premier temps, nous envisageons d'étudier les dataset publics ou proposés par divers services *open-data* comme BigQuery de GCP, Azure Open Datasets, AWS Open Data, data.gouv, etc. , pour des données structurées ou des plateformes comme Wikipedia et Imagenet pour les donnés non-structurées. En même temps, une collaboration sera mise en place avec un client de Zeenea, afin de valider les expérimentations sur un cas d'usage réel (catalogue de données anonymisé).

7 Organisation du travail

Le travail se fera principalement dans les locaux de Zeenea avec au minimum une journée par semaine au Laboratoire Informatique de Paris 6 (LIP6). Chez Zeenea, le candidat intégrera l'équipe R&D et sera au contact des équipes responsables du produit, ce qui lui permettra d'avoir une meilleure vision des défis et attentes sur le Smart Data Catalog. Au LIP6, le candidat intégrera l'équipe BD, qui compte des spécialistes du domaine de la data, et cela lui permettra d'être à jour sur les dernières technologies de pointe dans ce domaine.

8 Planning

Première année :

1. Étude de l'état de l'art scientifique
2. Proposer une modélisation du catalogue de données en graphe de connaissances
3. Définition d'une architecture physique pour le stockage des données
4. Définition de cas d'usage : données (open source, clients Zeenea), services d'enrichissement du graphe de connaissances (catalogue), définition de mesures de qualité
5. Publication d'un premier article sur le modèle de catalogue de données intelligent

Deuxième année :

1. Mise en place d'un catalogue de données test
2. Implémentation de nouvelles approches pour la suggestion de liens et l'enrichissement de propriétés : définition, implémentation, expérimentation et validation.
3. Publication de deux articles scientifiques, (i) Enrichissment des propriétés des objets du data catalog et (ii) détection de liens entre les objets du data catalog.

Troisième année :

1. Mise en place d'un prototype pour le Smart Data Catalog qui intègre les différents services intelligents proposés
2. Rédaction du manuscrit de thèse

9 L'entreprise Zeenea

Fondé en 2017, **Zeenea** est une jeune entreprise technologique, actrice de la French Tech, pour laquelle « la donnée doit être maîtrisée pour que les entreprises puissent innover ». Cette maîtrise est désormais fondamentale pour les entreprises, notamment avec l'émergence des innovations et de l'économie « data-centric ». L'importance des enjeux est indiscutable. L'information ne constituera une nouvelle richesse pour l'entreprise que si on sait l'exploiter et la valoriser. Une donnée peut être vue comme un actif numérique dès lors qu'elle comporte un potentiel de création de valeur pour l'entreprise. La question qui se pose est donc bien de savoir comment on optimise l'exploitation de ses données pour créer de la valeur.

Zeenea commercialise sa solution en mode SaaS uniquement et est déployé chez plus d'un trentaine de clients (banques, industries, transports, retails) dont la moitié à l'international. La société compte désormais 40 employés dont la moitié dans le département R&D.

10 Le laboratoire LIP6

Le **LIP6** est une Unité Mixte de Recherche (UMR 7606) de Sorbonne Université (SU) et du Centre National de la Recherche Scientifique (CNRS). Fort de plus de 500 membres, dont 200 permanents, il est l'un des plus grands laboratoires de recherche en informatique de France. Les 20 équipes du laboratoire couvrent un champ large des sciences informatiques : de l'électronique jusqu'à l'intelligence artificielle. Les collaborations du LIP6 relèvent tant de la recherche fondamentale (modélisation et résolution de problèmes fondamentaux) que de la recherche appliquée (mise en œuvre et validation de solutions en conditions réelles).

La thèse sera dirigée par **Bernd Amann** et co-encadré par **Camelia Constantin** et **Hubert Naacke**. Les trois encadrants font partie de l'équipe Bases de Données du LIP6 et ont une longue expérience de recherche dans la modélisation, l'interrogation et la gestion de données distribuées, volumineuses et complexes.

Sites web :

- LIP6 : https://www.lip6.fr
- Equipe BD : http://www-bd.lip6.fr/