Les entreprises utilisent un large éventail de systèmes d'information hétérogènes dans leurs activités commerciales telles que des systèmes pour la planification des ressources d'entreprise (ERP), la gestion des relations client (CRM) et la gestion de la chaîne logistique (SCM). En plus des grandes quantités de données hétérogènes produites par ces systèmes, les données externes sont une ressource importante pouvant être mise à profit pour permettre de prendre des décisions d'affaires rapides et rationnelles.

La Business Intelligence (BI) classique et même les nouveaux outils de visualisation Agile concentrent une grande partie de leurs caractéristiques de vente sur des visualisations attrayantes et uniques. La préparation des données pour ces visualisations reste une tâche beaucoup plus difficile dans la plupart des projets BI, qu'ils soient grands ou petits. Le provisionnement des données en libre-service vise à résoudre ce problème en fournissant la découverte intuitive des ensembles de données, ainsi que l'acquisition et l'intégration des données techniques pour l'utilisateur final.

L'objectif de cette thèse est de fournir un cadre qui permet l'approvisionnement des données en libre-service dans l'entreprise. Ce cadre permet aux utilisateurs métiers de rechercher, inspecter et réutiliser les données par le biais de profils de jeux de données sémantiquement enrichies.

Les ensembles de données publiquement disponibles contiennent des connaissances dans divers domaines tels qu'encyclopédique, gouvernemental, géographique, de divertissement et ainsi de suite. Il est difficile avec la diversité croissante de ces ensembles de données, de les annoter avec un nombre fixe de balises prédéfinies. En outre, les étiquettes saisies manuellement sont subjectives et ne peuvent pas capter leur essence et leur ampleur. Nous proposons un mécanisme pour attacher automatiquement des méta-informations aux objets des données en tirant parti des bases de connaissances comme DBpedia et Freebase ce qui facilite la recherche et l'acquisition de données pour les utilisateurs professionnels.

Dans de nombreuses bases de connaissances, des entités de données sont décrits avec de nombreuses propriétés. Cependant, toutes les propriétés ont la même importance. Certaines propriétés sont considérées comme clés pour effectuer des tâches d'instance tandis que d'autres propriétés sont généralement choisis pour fournir rapidement un résumé des faits principaux attachés à une entité.

Les utilisateurs professionnels peuvent vouloir enrichir leurs rapports avec ces entités de données. Pour faciliter cela, nous proposons un mécanisme pour sélectionner quelles propriétés devraient être utilisés lors de l'extension des colonnes supplémentaires dans un jeu de données existant où annoter des cas avec des balises sémantiques.

Le principe de données ouvertes liées (LOD) a émergé comme étant l'une des plus grandes collections d'ensembles de données interdépendantes sur le Web. Afin de bénéficier de cette mine de données, on a besoin d'accéder à des informations descriptives sur chaque jeu de données (ou métadonnées). Ces métadonnées permettent la découverte de données, leur compréhension, leur intégration et leur maintenance. Les portails de données, qui sont les points d'accès de ces jeux de données, offrent des métadonnées représentées dans des modèles différents et hétérogènes. Nous proposons d'abord un

modèle de données harmonisé basé sur une enquête systématique de la littérature permettant une couverture complète des métadonnées afin de permettre la découverte de nouvelles données, leur exploration et leur réutilisation par les utilisateurs professionnels. Deuxièmement, des métadonnées riche d'informations sont actuellement très limitée à quelques portails de données où elles sont généralement fournies manuellement, étant donc souvent incomplètes et incohérentes en termes de qualité. Nous proposons une approche évolutive automatique pour extraire, valider, corriger et générer des profils d'ensembles de données liées descriptives. Cette approche applique plusieurs techniques afin de vérifier la validité des métadonnées fournies et pour générer des informations descriptives et statistiques pour un ensemble de données particulier ou pour un portail de données entier.

La qualité des données est un domaine traditionnel de recherche approfondie avec plusieurs points de repère et des cadres pour saisir ses dimensions. Assurer la qualité des données dans le principe de données ouvertes liées est beaucoup plus complexe. Elle se compose de l'information structurée soutenue par des modèles, des ontologies, des vocabulaires et contient des liens terminaux et interrogeables. Nous proposons un cadre d'évaluation objective de la qualité des données liées basés sur des métriques de qualité pouvant être mesurée automatiquement. Nous présentons en outre un outil extensible de mesure de la qualité par la mise en œuvre de ce cadre afin d’aider les propriétaires de données d’évaluer à la main la qualité de leurs jeux de données, d’obtenir des conseils sur les améliorations possibles et d’aider les autres consommateurs de données afin de choisir leurs sources de données à partir d'un ensemble classé.

Enfin, l'Internet a créé un changement de paradigme dans la façon dont nous consommons et diffusons l'information. Les données actuelles sont réparties sur des silos hétérogènes de données archivées et en direct. Les gens partagent volontiers les données sur les médias sociaux en affichant des nouvelles, des vues, des présentations, des photos et des vidéos. Nous proposons un service qui apporte de la pertinence en direct, et archive l’information pour l'utilisateur d'affaires. Le principal avantage est un accès instantané à des informations complémentaires sans avoir à chercher. Les informations apparaissent quand elles sont pertinentes permettant à l'utilisateur de se concentrer sur ce qui est vraiment important.