Interrogation du web sémantique

Jérôme David, Jérôme Euzenat, Nabil Layaïda, Nabil Layaïda, Marie-Christine Rousset

Du Big Data à l'IA - 60 ans d'expériences en traitement des Données, des Informations et des Connaissances à Grenoble

sous la direction de Michel Adiba et Jean-Pierre Giraudin

Chapitre IV-3 **: Interrogation du Web sémantique**[*]

Rédacteurs : Jérôme David[a], Jérôme Euzenat[a], Nabil Layaïda[a], Marie-Christine Rousset[b]

Avant-propos

Le Web est un énorme gisement d'informations distribuées et hétérogènes où cohabitent des formats et des ressources très variés. Les pages Web, identifiées par une adresse sur Internet, constituent un graphe de documents dans lequel on peut chercher de l'information par mots-clefs grâce à un **moteur de recherche** comme Google. L'adoption par le W3C au début des années 2000 du modèle de données RDF (Resource Description Framework) comme langage standard pour associer des métadonnées à n'importe quelle ressource du Web a permis l'avènement du **Web des données** que l'on peut interroger à l'aide du **langage de requêtes SPARQL** (cf. chapitre I-3). Le **Web sémantique** est un enrichissement du Web des données par des **ontologies,** qui sont des connaissances permettant d'associer une sémantique claire aux ressources du Web et à leurs métadonnées. Le Web sémantique est donc un immense graphe **de connaissances** où des algorithmes de raisonnement automatique permettent de renforcer les langages de requêtes et ainsi enrichir l'information qu'on peut trouver sur le Web.

Le web sémantique se fonde sur des sources de données décrites sémantiquement, distribuées et indépendantes. Son exploitation ambitionne de répondre à des requêtes en raisonnant sur ces sources. Il pose donc de nouveaux problèmes à l'exploitation des données et des connaissances : disposer de langages adaptés à l'expression des connaissances sur le web, cela concerne les données, ontologies, alignements et requêtes ; disposer d'algorithme efficaces pour les manipuler, en particulier vérifier la cohérence entre sources, évaluer des requêtes exploitant les ontologies, trouver des correspondances entre ontologies, ou reconnaître des descriptions redondantes dans des sources distribuées. La recherche grenobloise a amplement contribué à la résolution de certains d'entre eux : raisonner en présence d'ontologies, trouver et exprimer les correspondances entre ontologies, trouver et raisonner avec les entités dupliquées et exprimer et évaluer efficacement des requêtes.

La recherche grenobloise sur l'interrogation de graphes de données et sur les ontologies comme interface de requêtes sur des données hétérogènes est très visible sur le plan national et international et s'appuie sur plusieurs équipes universitaires et de l'Inria avec des compétences complémentaires en base de données et représentation de connaissances.

1 Raisonner sur des données en présence d'ontologies

RDF (*Resource Description Framework*) est le standard défini par le consortium W3C de standardisation du Web pour associer des métadonnées à n'importe quelle ressource référencée sur le Web par un URI (*Uniform Resource Identifier*) en déclarant des propriétés sur ces URIs dans une base de métadonnées constituée de triplets RDF et interrogeable à l'aide du langage de requêtes SPARQL.

Prenons à titre d'illustration le site de DBpedia qui est un projet universitaire et communautaire d'exploration et d'extraction automatiques de données dérivées de Wikipédia. Son principe est de proposer une version structurée et normalisée au format du web sémantique des contenus de Wikipedia. Par exemple les deux URI ci-dessous identifient respectivement Paris (http://fr.dbpedia.org/page/Paris) et l'acteur et cinéaste François Ozon

[*] Version auteurs qui diffère substantiellement de la version publiée.

[a] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France

[b] Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble, France

(http://fr.dbpedia.org/page/François_Ozon). Ainsi le triplet RDF : (http://fr.dbpedia.org/page/Paris, http://fr.dbpedia.org/ontology/birthPlace, http://fr.dbpedia.org/page/François_Ozon) exprime que la ville de Paris est le lieu de naissance de François Ozon en reliant ces deux entités par la propriété « avoir pour lieu de naissance » identifiée elle-même par l'URI http://fr.dbpedia.org/ontology/birthPlace.

Contrairement aux vocabulaires de métadonnées comme Dublin Core[1], restreints à un ensemble fermé de termes prédéfinis, RDF permet de combiner dans une même base différents vocabulaires de métadonnées (provenant d'espaces de noms différents) et de définir ses propres métadonnées qui peuvent être à leur tour réutilisées par d'autres. Cette flexibilité et cette expressivité ont permis le déploiement rapide du Web des données (*Linked Data*). Aujourd'hui, le Web des données (Figure 1) regroupe plusieurs centaines de milliards de triplets RDFs répartis dans plusieurs milliers de bases de données RDF, comme DBpedia.fr, accessibles et interrogeables par des points d'entrée SPARQL sur le Web.

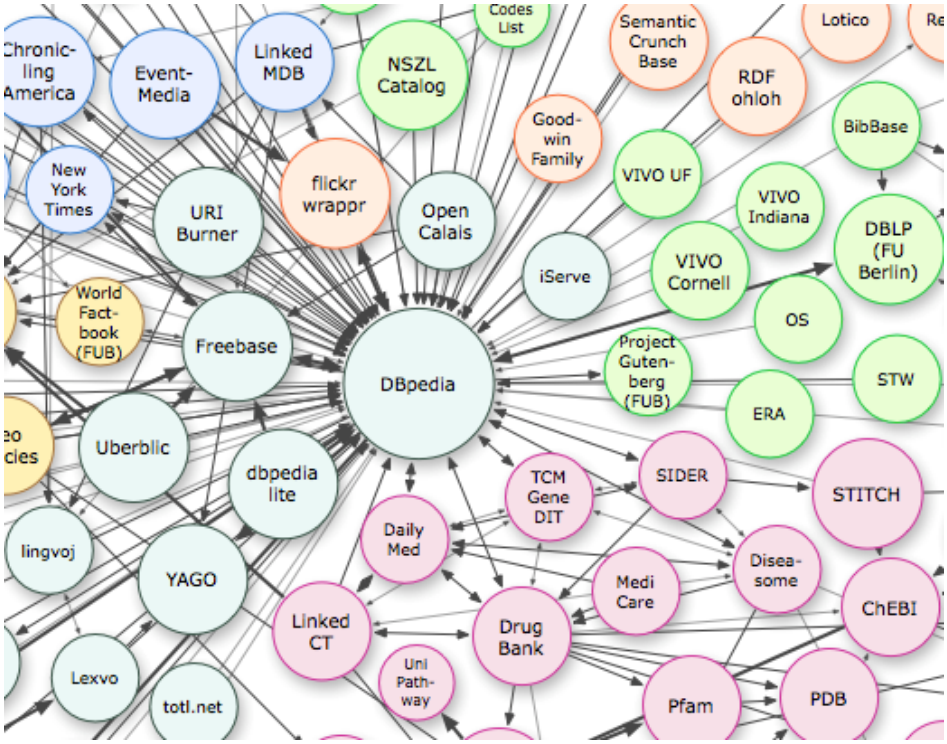

Figure 1. Extrait du Web des données centré sur DBpedia qui sert de référentiel pour des sources de données bibliographiques (en vert) mais aussi des bases de données du domaine biomédical (en rose).

Le traitement automatique de métadonnées à l'aide d'**ontologies** est la clé de voûte du **Web sémantique**. Les ontologies sont des connaissances formalisées, partagées par une communauté, sur les classes/concepts et les propriétés/relations d'un domaine d'application. Les logiques de description ont fourni le cadre formel de la définition du langage OWL (*Ontology Web Language*), le standard du W3C pour représenter les ontologies du Web sémantique. Certaines ontologies « légères » peuvent être exprimées en RDF. Par exemple, la connaissance que la classe des acteurs est une sous-classe de celle des artistes peut être déclarée par le triplet RDF suivant : (http://dbpedia.org/ontology/Actor, rdfs:subClassOf, http://dbpedia.org/ontology/Artist). Ce triplet doit être interprété comme une règle générale permettant d'inférer pour tout *x* que si *x* est un acteur alors *x* est un artiste. Interroger des données décrites par des ontologies devient un problème de raisonnement et des algorithmes d'inférence sont nécessaires pour pouvoir inférer l'ensemble des réponses à une requête, même simple. Ainsi, en

[1] Le Dublin Core a été développé par la Dublin Core Metadata Initiative, aussi connue sous le sigle DCMI, pour décrire des documents de manière simple et standardisée à l'aide d'un ensemble prédéfini de 10 propriétés.

l'absence d'algorithme d'inférence permettant d'activer la règle ci-dessus, François Ozon, déclaré comme instance de « Actor », ne sera pas trouvé comme réponse à la requête mot-clé « Artist ».

De nombreux domaines se dotent progressivement d'ontologies métiers pouvant être interconnectées dans le Web des données ou servir de médiateurs pour l'intégration sémantique d'informations hétérogènes dans des applications variées.

Les ontologies fournissent une vue conceptuelle des données et services mis à disposition sur le Web, dans le but de faciliter leur manipulation. Répondre à des requêtes sur des ontologies est un problème central pour la mise en œuvre du Web sémantique. Une étape de reformulation de la requête en fonction des axiomes et contraintes déclarés dans l'ontologie est nécessaire pour garantir la complétude des réponses. Le point important est que cette reformulation (comme la réécriture de requêtes en termes de vues dans le domaine des bases de données) est un problème de raisonnement sur la requête et les axiomes de l'ontologie qui est indépendant des données. L'étude de ce nouveau problème a donné lieu à une communauté très active de chercheurs à la croisée de la représentation de connaissances et les bases de données. Différents langages logiques à base de règles ou de logiques de description ont été proposés et comparés en termes d'expressivité et de complexité du raisonnement associé. Ces travaux s'inscrivent dans le domaine de recherche émergent d'Ontology-Based Data Access pour lequel des algorithmes efficaces ont été proposés pour l'interrogation, l'intégration, l'analyse et le liage de données hétérogènes et de qualité variable. Certains de ces algorithmes sont décrits dans l'ouvrage de référence « Web Data Management » dont Marie-Christine Rousset est co-auteur.

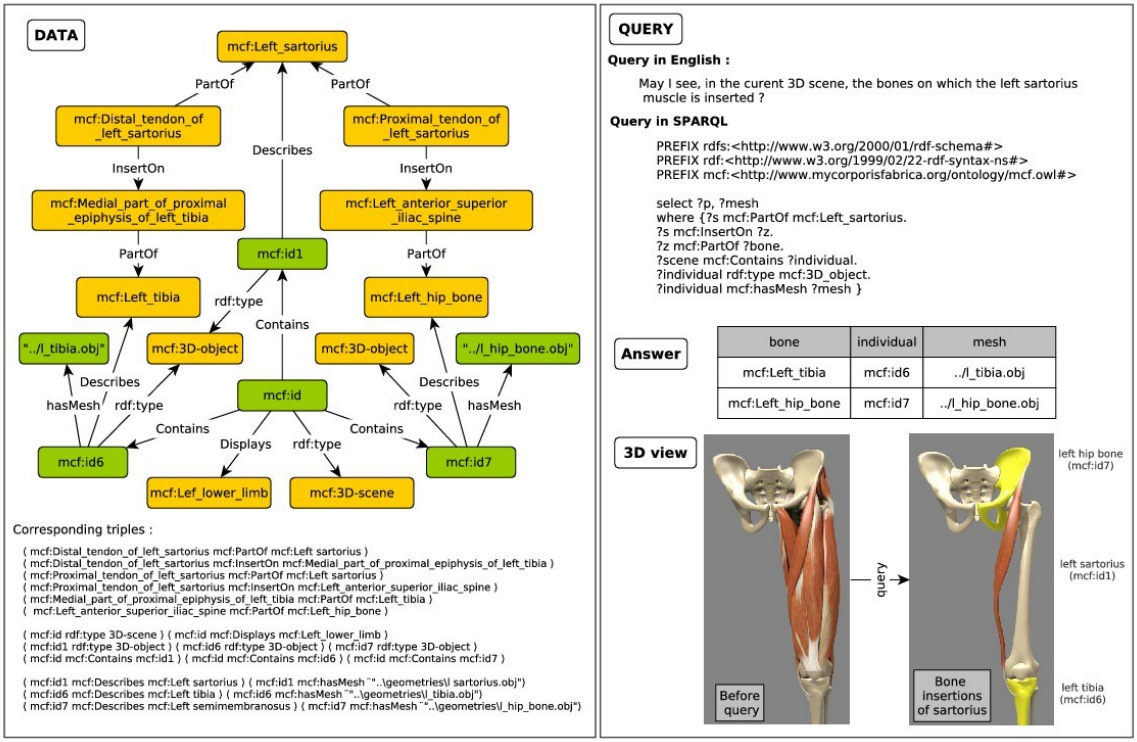

Figure 2. Extrait de l'ontologie My Corporis Fabrica, développée avec Olivier Palombi

L'équipe SLIDE du LIG a été impliquée dans plusieurs projets liés à cette nouvelle problématique, en particulier dans le domaine de l'enseignement de la médecine. La figure 2 illustre l'approche que nous avons développée en collaboration avec Olivier Palombi, professeur d'anatomie à l'Université Grenoble Alpes, pour intégrer dans une même ontologie différents types de connaissances liées à l'anatomie pour une visualisation 3D à la demande.

Ce travail a été repris et enrichi et l'ontologie My Corporis Fabrica est au coeur de la technologie "Digital Anatomy for Personalized Healthcare" développée par la startup Anatoscope[2].

Toujours en collaboration avec Olivier Palombi, devenu vice-président de l'UNESS (Université Numérique en Santé et Sport), nous avons été fortement impliqués dans le projet SIDES 3.0 financé par l'ANR dans le cadre du programme d'investissement d'avenir DUNE pour le développement des universités numériques expérimentales. SIDES 3.0 a permis de mettre en place les briques de base de la plateforme d'apprentissage personnalisée en médecine SIDES NG développée par l'UNESS et utilisée par toutes les facultés de médecine de France. Le cœur de SIDES 3.0 est le graphe de connaissances OntoSIDES résultant de l'intégration par une approche à base d'ontologies et de mappings, de données de traces d'apprentissage de plus de 140 000 étudiants en Médecine sur une période de près de 10 ans. L'exploitation du graphe de connaissances OntoSIDES a donné lieu à une première analyse des résultats de l'apprentissage des étudiants qui va permettre la conception et la mise en œuvre de fonctionnalités d'auto-entraînement personnalisé et adaptatif. Le projet SIDES LAB, financé par l'ANR pour 4 ans à partir de 2022, exploite les résultats de SIDES 3.0 pour conduire des expérimentations in situ en sciences de l'éducation. La méthodologie que nous avons suivie n'est pas spécifique à l'enseignement de la Médecine. Elle peut être appliquée à d'autres disciplines, à condition qu'un référentiel partagé par la communauté existe pour définir de manière précise et consensuelle le programme et les objectifs pédagogiques sous-jacents aux questionnaires d'évaluation ou d'entraînement pour cette discipline.

Nous avons contribué à fédérer la communauté française sur l'intégration de données par des ontologies (Ontology-based Data Access) au travers de notre participation aux projets ANR PAGODA (Practical Algorithms for Ontology-based Data Access) et CQFD (Complex ontological Queries over Federated and heterogeneous Data) , ainsi que par la création et l'animation du groupe de travail RoD (Reasoning on Data) commun aux GDR MADICS et IA du CNRS, qui a débouché en 2021 sur l'action RoCED (Reasoning on Complex and Evolving Data).

2 Hétérogénéité sémantique et alignement d'ontologies [2000-2022]

En 2000, l'équipe Exmo[3] fut créée pour introduire la représentation de connaissance au cœur du web (et confirmée équipe-projet INRIA en 2003). C'était l'un des prolongements du projet INRIA Sherpa (voir Chapitre II-4) qui avait depuis longtemps déjà expérimenté l'édition collaborative de bases de connaissance sur le web[4] (Hytropes, Co4). Sherpa avait aussi coordonné l'action RNTL Escrire impliquant les équipes INRIA Sherpa, Orpailleur (Nancy) et Acacia (Sophia-Antipolis) destinée à déterminer les mérites et défauts de différentes représentations de connaissances pour le Web[5]. Pour cette raison, Exmo s'est rapidement impliqué dans la construction du web sémantique. Jérôme Euzenat a coorganisé le *Semantic web working symposium* à Stanford en 2001 qui deviendra la série d'*International semantic web conference*[6] (ISWC). Il a aussi organisé la même année à Sophia-Antipolis une rencontre EU-NSF sur le sujet[7]. L'équipe a participé activement au groupe de travail du W3C qui a abouti à la proposition de recommandation de la première version du langage d'ontologies OWL ainsi qu'au groupe de travail sur le langage de requête SPARQL.

Le réseau d'excellence *Knowledge web*[8] créé en 2004, fut entre autres choses, l'occasion de structurer l'activité autour de l'alignement d'ontologies (ontology matching). Il s'agit d'établir les relations entre deux ontologies et de pouvoir les exploiter. Ces relations peuvent être l'équivalence entre deux classes ou propriétés. *Livre* = *Book* (livre et book sont la même chose). Il peut aussi s'agir de relations plus complexes, comme: *Biographie* $\sqsubseteq$ *Book* (une biographie est une sorte de livre) or *Biography* = *Book* $\sqcap$ $\exists$*topic.Person* (une biographie est un livre dont un sujet est une personne). L'avantage de disposer d'une notion d'alignement indépendante et pouvant s'exprimer de manière *déclarative*, en particulier indépendamment de son usage, est de pouvoir exploiter les alignements dans de

[2] https://www.anatoscope.com/
[3] https://exmo.inria.fr
[4] https://co4.inrialpes.fr
[5] https://escrire.inrialpes.fr
[6] http://swsa.semanticweb.org/content/international-semantic-web-conference-iswc
[7] https://www.ercim.eu/EU-NSF/semweb.html
[8] http://knowledgeweb.semanticweb.org

nombreuses tâches: transformer des assertions (données, formules, requêtes, messages entre agents) utilisant une ontologie dans une autre, fusionner ou rapprocher deux ontologies, permettre de négocier la communication entre différents services et agents. Certaines de ces applications sont relativement statiques se satisfaisant d'un alignement hors-ligne alors que d'autres, très dynamiques, nécessitent des capacités d'alignement embarquées permettant de faire évoluer les alignements au cours du temps.

Des travaux similaires avaient déjà été menés dans le cadre de la mise en correspondance de schémas de bases de données afin de pouvoir importer des données d'une base à l'autre. Mais les représentations de connaissance formelles, comme OWL, rendent le problème plus difficile à cause de leur expressivité accrue. En contrepartie, utiliser des langages définis sémantiquement permet de mieux évaluer la qualité des alignements.

Les travaux menés sur ce sujet ont été très riches et ont abouti à de nombreux résultats de différentes natures:

- Un langage simple pour exprimer les alignements, qui reste le standard de fait, a été défini. Il a été étendu à l'expression d'alignements plus complexes permettant de poser des contraintes sur les entités liées par l'alignement et d'exprimer des transformations (EDOAL). Ce langage d'alignement a permis d'exprimer des réseaux d'ontologies, notion qui sera approfondie dans un autre projet Européen (NeOn pour *Networked ontologies*).
- Différentes sémantiques ont été associées à ce langage. Disposer d'une sémantique permet de développer des travaux fondés sur des bases solides. C'est ainsi que l'équipe Exmo a contribué au raisonnement avec des alignements et des réseaux d'ontologies, la composition d'alignements utilisant le formalisme des algèbres de relations, la clôture de la connaissance exprimée et la révision de la connaissance ainsi distribuée lorsqu'elle vient à être contradictoire.
- L'équipe a aussi défini une interface fonctionnelle (API) permettant de manipuler les alignements et les réseaux d'ontologies. Elle a aussi produit une implémentation de cette interface fonctionnelle qui a été utilisée par plus de 50 systèmes dans le monde pour lequel un article a été publié.
- Knowledge web, et Exmo au premier chef, est à l'origine d'une méthodologie d'évaluation des alignements, et des systèmes les produisant. D'un point de vue théorique, différentes mesures, inspirées de la précision et du rappel, ont été développées pour comparer les alignements. Toutes ces techniques ont été mises en œuvre dans la campagne annuelle d'*Ontology alignment evaluation initiative*[9] (OAEI) qui consiste à évaluer les algorithmes d'alignements suivant diverses modalités proposées par les organisateurs. Créée en 2005, OAEI existe toujours (elle a aussi inspiré un projet Européen, SEALS pour *Semantic evaluation at large scale*).
- Plusieurs techniques d'alignement ont été développées (recherche d'appariement de similarité maximale, alignement dirigée par les contextes, composition d'alignement) et implémentées dans des systèmes: OLA, Aroma (développé par Jérôme David avant d'arriver à Grenoble).
- Enfin, différentes mesures pour comparer les éléments des ontologies et ainsi aider à la détermination des alignements ont été développées. Elles peuvent être fondées sur des indices aussi variés que le lexique utilisé, la structure des objets, la quantité d'information qu'elles apportent, la proportion d'objets communs qu'elles dénotent, leur consistance mutuelle ou une combinaison de tels facteurs.
- Finalement, les ateliers annuels sur l'alignement d'ontologies, créés en 2006 dans ce même cadre, sont toujours très fréquentés en 2022.

Ces travaux ont été en grande partie dirigés et produits à Grenoble en collaboration avec de nombreux collègues participants aux projets européens OntoWeb, Knowledge web, NéOn, SEALS, Ready4SmartCities et français WebContent, WebIntelligence et DataRing (et toutes les propositions de projets qui ont échoué mais ont enrichi la réflexion et renforcé les liens entre les participants). Les activités menées au sein de Knowledge web sont à l'origine de l'ouvrage de référence *Ontology matching* rédigé par Jérôme Euzenat et Pavel Shvaiko en 2007.

Lors de la publication de la deuxième édition de l'ouvrage *Ontology matching*, en 2013, l'équipe Exmo a réfléchi à ce qui pouvait être fait pour compléter le travail sur l'alignement d'ontologies. Cela a conduit à deux lignes de travaux : le

[9] https://oaei.ontologymatching.org

liage des données et l'étude de l'évolution de la connaissance. Ces deux directions ont été poursuivies dans l'équipe mOeX. On se concentre ci-dessous sur le lien le plus direct : le liage de données.

3 Liage de données [2010-2022]

En 2006, suite au constat que le travail sur le web sémantique n'avait pas un impact massif à cause du manque de données, Tim Berners-Lee a proposé l'idée de données liées: l'expression de jeux de données en RDF, accessibles depuis le web. La manière standard de lier ces données consiste à publier des liens « sameAs » entre ressources décrites dans différents jeux de données, signifiant qu'elles dénotent la même ressource. Ce sont ces liens qui forment les connexions entre les nœuds de la Figure 1. Déterminer les liens entre diverses sources de données indépendantes est particulièrement ardu, c'est ce que nous appelons le liage de données.

Il y a deux grandes familles de techniques pour effectuer cette opération le plus automatiquement possible :

- définir une mesure de similarité entre les descriptions RDF et considérer qu'une similarité élevée correspond à la même ressource;
- déterminer des conditions nécessaires pour que deux descriptions correspondent à la même ressource et appliquer ce critère.

Le site grenoblois (équipe Exmo puis Moex, HADAS puis SLIDE) en collaboration avec des collègues d'Orsay (LRI) et Montpellier (LIRMM) est un pilier de ce que l'on peut qualifier d'école française de liage de données, plus focalisée sur la seconde approche.

En 2010, l'équipe Exmo a été à l'initiative du projet ANR DataLift qui visait à favoriser l'émergence du Web des données en développant une plateforme pour structurer, publier et interconnecter des jeux de données sur le web de données. L'approche suivie par l'équipe pour le liage de données s'appuie principalement sur la notion de dépendance fonctionnelle et plus précisément sur l'extraction de clés. Une clé est un ensemble de propriétés permettant de discriminer l'ensemble des individus d'un graphe ou d'une classe. Comme les données sont généralement bruitées, l'équipe a introduit la notion de pseudo-clé qui autorise quelques exceptions. Un algorithme efficace d'extraction de pseudo-clés à partir de graphes RDF a été développé et testé sur des jeux de données volumineux comme DBpedia. Les résultats montrent son utilité dans des optiques de détection de redondance et/ou d'erreurs et de liage de jeux de données RDF.

La sémantique des clés utilisée dans ces travaux diffère de celles utilisées classiquement en base de données et dans le langage OWL. Une comparaison théorique et expérimentale des différentes sémantiques de clés proposées a été réalisée dans le cadre du projet ANR Qualinca, en collaboration avec des équipes du LRI (Orsay) et du LIRMM (Montpellier). Il a été entre-autre montré que la notion de clé utilisée dans les travaux sur les pseudo-clés généralise celle de OWL tout en permettant de couvrir des clés qui sont intéressantes en pratique.

Toujours dans le cadre du projet Qualinca, et plus particulièrement d'une collaboration entre les équipes HADAS et Exmo, une méthode de liage de données basée sur l'utilisation d'un Datalog probabiliste a été développée. Les résultats montrent que (1) l'approche passe à l'échelle, (2) le raisonnement permet de lier des données qui ne l'auraient pas été en utilisant des outils à l'état de l'art de l'époque, (3) que la prise en compte de règles incertaines permet d'augmenter grandement le rappel tout en conservant une bonne précision.

Dans le cadre du projet ANR-Blanc Lindicle associant Exmo et une équipe de l'université de Tsinghua (Pékin), le liage de données multilingues a été également étudié. Contrairement aux précédents travaux, l'approche suivie ne se base pas sur la notion de clé, mais représente des entités RDF par des documents virtuels et les compare suivant les métriques classiques de recherche d'information. Afin de traiter l'aspect multilingue, deux directions ont été considérées : (1) la traduction automatique des documents virtuels ; (2) une approche indépendante du langage où les termes importants sont remplacés par leur sens (extrait d'une ressource terminologique comme BabelNet,

Wikipedia). Il a été montré que cette approche fonctionne bien pour le liage d'entités nommées et également sur des thésauri.

Depuis 2011, le travail sur les clés a été étendu à la notion de clé de liage (*link key*). Une clé de liage est une assertion qui permet d'identifier les instances communes à deux jeux de données. Un exemple de clé de liage est: {⟨auteur, creator⟩}{⟨titre, title⟩} linkkey ⟨Livre, Book⟩ indiquant que si une instance de la classe Livre a les mêmes valeurs pour la propriété auteur qu'une instance de la classe Book a pour la propriété creator et qu'elles partagent au moins une valeur de leurs propriétés titre et title, alors elles dénotent la même ressource. Contrairement aux clés classiques, les clés de liage sont définies entre deux ontologies (comme les alignements). Il a été montré qu'une clé de liage est strictement plus générale d'une paire de clés dont les propriétés sont liées par un alignement. L'équipe Exmo a développé un algorithme de découverte de clés de liage à partir de données et proposé des mesures originales permettant d'évaluer la qualité des règles extraites dans les cas supervisé et non supervisé. Les résultats montrent que ces deux familles de mesures sont de bonnes approximations de la précision et du rappel et que l'algorithme développé est particulièrement robuste.

Suite à ce premier travail sur l'extraction de clés de liage, des liens avec les techniques de l'analyse formelle de concepts (FCA) ont été établis. Il a été notamment proposé une méthode d'extraction de clés de liage interdépendantes en utilisant l'analyse relationnelle de concepts (RCA). La méthode permet de trouver des clés de liages qui s'appuient sur les liens engendrés par d'autres clés de liages, par exemple à trouver une clé de liage pour identifier des livres en s'appuyant sur leur titre et auteurs qui sont eux-mêmes identifiés par une clé de liage reposant par exemple sur le nom et les prénoms, voire sur les ouvrages qu'ils ont écrits. L'extraction de disjonctions de clés de liage a été également étudiée. L'objectif est d'améliorer le rappel tout en gardant une bonne précision. Étant donné que le nombre de disjonctions est exponentiel en fonction du nombre de propriétés, il n'est pas envisageable de toutes les énumérer. Par conséquent, deux heuristiques pour extraire les « meilleures » disjonctions ont été proposées et étudiées expérimentalement.

Le projet ANR Elker, en collaboration avec l'INRIA Lorraine (Nancy) et L'université de Paris 8, a permis de poursuivre les travaux sur les clés de liage selon deux directions. D'une part, une technique d'extraction de clés de liage tirant parti du formalisme des « Pattern Structures » a été proposée. Un tel algorithme permet de calculer la paire d'expressions de classes associée à une clé de liage candidate. Un autre sujet est le raisonnement avec les clés de liage. En effet, ces clés peuvent être interprétées comme des axiomes de logiques de descriptions. Ainsi, avec deux ontologies et des clés de liage entre elles, il est bien sûr possible de déduire les liens, mais aussi d'autres clés de liage et d'autres axiomes. Il est de surcroît possible de déterminer si l'ajout de telles clés entre deux ontologies ne les rend pas inconsistantes. Afin de rendre effectif ce raisonnement, un mécanisme d'inférence a été défini et implémenté en étendant plusieurs raisonneurs de logiques de descriptions.

4 Extension et optimisation de requêtes SPARQL

Les bases de données organisées en graphes sont l'une des extensions proposées pour le Big Data. Dans ce contexte, l'un des challenges est de définir et mettre en œuvre des langages de requêtes, qui à cause de la nature des graphes, doivent supporter la récursion. Plusieurs extensions ont été proposées pour exprimer des requêtes récursives, que ce soit pour SPARQL (CPSPARQL) ou pour l'algèbre relationnelle (muRA). C'est dans ce contexte que se placent les travaux de l'équipe Tyrex, et parfois Exmo, de l'INRIA et du LIG. Ces travaux ont porté principalement sur (1) l'extension de SPARQL aux expressions de chemins, (2) l'analyse statique de requêtes, (3) l'optimisation de requêtes récursive, (4) l'évaluation distribuée de muRA et (5) l'évaluation à très grande échelle des requêtes SPARQL.

4.1. Extension de SPARQL aux expressions de chemin

Nous avons défini CPSPARQL, étendant SPARQL de telle sorte que les prédicats des triplets puissent être remplacés par des chemins exprimés par des expressions régulières supportant des contraintes. Par exemple, il est possible de chercher les connexions en train ou bus entre deux villes tant qu'elles offrent le Wifi. Cette extension a été reprise, sous une forme altérée dans le standard SPARQL 2. Nous avons développé un algorithme complet et correct pour calculer leurs projection dans les graphes RDF. Cela permet d'évaluer les requêtes CPSPARQL. Bien que ce problème

reste NP-dificile, il est possible de développer une évaluateur robuste et efficace. Enfin, nous avons montré comment réécrire des requêtes dépendant d'une ontologie exprimée dans un fragment de RDFS, en une requête CPSPARQL équivalente.

4.2. Analyse statique de requêtes SPARQL

Le travail sur l'analyse statique a porté sur l'étude du problème de l'inclusion de requête SPARQL. Ce problème permet de déterminer si le résultat d'une requête est inclus dans le résultat d'une autre requête pour tout graphe RDF. Il a des applications importantes dans l'optimisation des requêtes et la vérification de bases de connaissances. Nous avons développé une procédure de décision afin de déterminer l'inclusion des requêtes sur des jeux de données obéissant à des schémas décrits sous forme d'axiomes en logique de description. Plusieurs techniques de test d'inclusion ont été proposées dans l'état de l'art en utilisant différentes techniques : homomorphisme de graphes, bases de données canoniques, les techniques de la théorie des automates.

Nous avons choisi de traiter ce problème par une réduction au problème de la validité de formules modales de graphes en logique. En particulier, nous avons utilisé une logique expressive appelée µ-calcul équipée des modalités inverses (permettant d'exprimer des relations inverses) ainsi que deux point fixes pour supporter les différentes formes de récursion. Les données RDF sont représentées sous formes de systèmes de transitions qu'on retrouve communément dans le domaine de la vérification. Les requêtes et les axiomes d'un schéma sont représentés sous forme de formules qui contraignent les modèles possibles et donc les jeux de données RDF. Le problème de l'inclusion de requêtes peut alors être réduit à un test de l'invalidité de formules logiques. On cherche en pratique à trouver un graphe qui invalide la formule et si on n'en trouve pas, cela prouve que la formule est valide. Dans notre cas, on procède par énumération efficace des modèles en utilisant des diagrammes de décision binaire de ces modèles.

Nous avons ensuite cherché à caractériser divers fragments de SPARQL (et PSPARQL) et langages de description logique de schéma pour lesquels le test d'inclusion est décidable. En particulier, nous avons isolé un fragment qui dispose de la propriété de modèle d'arbre (Tree Model Property). Cette propriété est intéressante car si elle est satisfaite, pour rechercher un graphe contre-exemple d'une formule il suffit de rechercher un arbre contre-exemple.

Afin de démontrer la viabilité pratique de cette méthode, nous avons développé un solveur logique dédié. Nous avons également testé ce solveur sur un nouveau benchmark construit à partir de requêtes réelles observées sur le jeu de données Wikipédia[10].

4.3. Optimisation et évaluation efficace de requêtes SPARQL

Le deuxième axe de recherche a porté sur l'optimisation et l'évaluation efficace des requêtes SPARQL. La contribution principale de l'équipe est l'élaboration d'une nouvelle méthode particulièrement adaptée pour des requêtes contenant de la récursion ou destinée à une évaluation distribuée. Elle s'appuie sur une généralisation de l'algèbre relationnelle appelée muRA et publiée dans la conférence SIGMOD 2020. Cette algèbre est équipée d'un opérateur de point fixe qui permet d'exprimer de nouvelles formes de récursion. Nous avons défini une syntaxe et une sémantique pour muRA ainsi qu'une traduction complète des requêtes SPARQL avec des expressions de chemins : Property Paths. Il s'agit d'une fonctionnalité introduite dans le standard qui permet de sélectionner des nœuds accessibles par des chemins récursifs.

Nous avons également défini un système de types pour muRA et avons montré comment les termes de muRA peuvent être réécrits en d'autres termes ayant une sémantique équivalente en utilisant soit des règles de réécriture provenant de l'algèbre relationnelle classique soit des règles nouvelles, spécifiques à cette nouvelle algèbre. Nous avons démontré la correction des nouvelles règles qui sont introduites pour réécrire les points fixes : elles permettent de pousser les filtres, les jointures ou les projections à l'intérieur des points fixes en fonction de certaines conditions sur les termes. Ces réécritures particulières permettent d'exprimer des optimisations avec de nouveaux plans (c.f. figure 3) qui ne peuvent pas être réalisées par des systèmes comme Datalog.

[10] https://sparql-qc-bench.inrialpes.fr

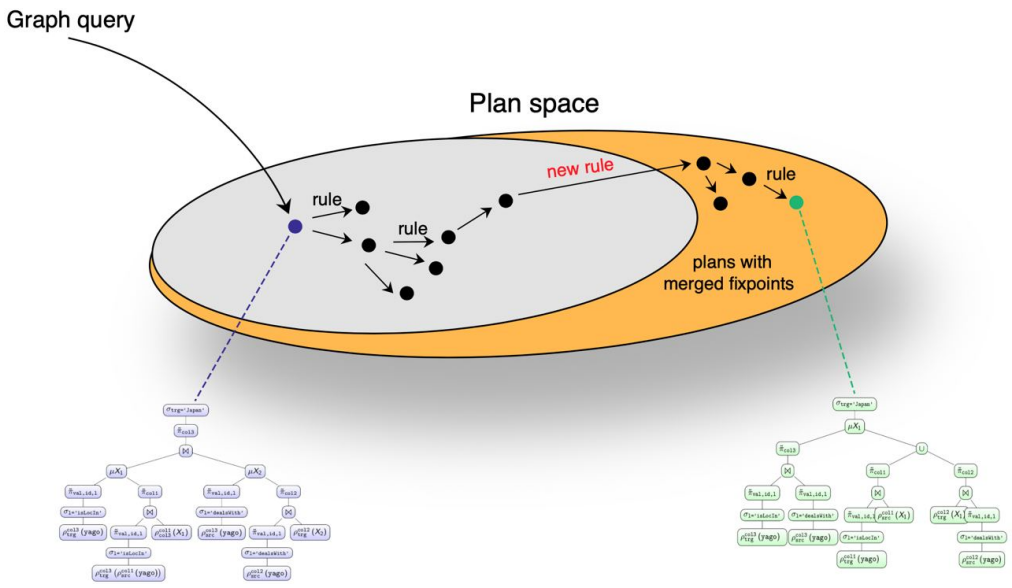

Figure 3 : Optimisation des requêtes

Nous avons ensuite étudié comment ces termes peuvent être évalués, d'abord de manière générale, puis en considérant une évaluation sur une plateforme distribuée. Pour cela, nous avons développé un modèle de coût pour l'évaluation des termes. À l'aide du modèle de coût et de l'évaluateur, plusieurs termes qui sont équivalents d'un point de vue sémantique peuvent maintenant être vus comme différentes manières d'évaluer les termes avec différents coûts estimés. Nous avons alors montré que les termes qui sont considérés grâce aux nouvelles règles de réécritures, permettent une exécution plus efficace que ce qui était possible avec les approches existantes comme Datalog ou Postgres. Ils correspondent à des parcours plus variés du graphe par exemple du milieu d'un chemin vers ses extrémités ou encore de la fin d'un chemin vers son début, et qui peuvent se révéler plus efficaces compte tenu des cardinalités des nœuds impliqués dans ces parcours.

4.4. Évaluation distribuée de muRA

DistMuRA est un système qui évalue l'algèbre muRa de façon distribuée sur Spark[11]. Différentes stratégies d'évaluation de termes algébriques récursifs dans un contexte distribué ont été étudiées. Ces stratégies sont implémentées sous forme de plans physiques avec des techniques qui automatisent la distribution des données afin de réduire les coûts de communication. Les résultats expérimentaux sur des graphes réels et synthétiques montrent l'efficacité de l'approche proposée par rapport aux systèmes existants.

4.5. Évaluation distribuée efficace de requêtes SPARQL

Nous avons complété ce résultat théorique par une expérimentation comparant plusieurs exécuteurs sur des requêtes SPARQL contenant de la récursion. Ensuite, nous avons conçu un évaluateur qui s'appuie sur un fragment de muRA restreint aux opérateurs qui ont une traduction en code Spark efficace. Le résultat de ce travail s'est traduit par l'implémentation de SPARQLGX, un évaluateur SPARQL distribué en pointe par rapport à l'état de l'art. Pour finir, la dernière contribution concerne l'estimation de la cardinalité des solutions à un terme de la µ-algèbre. Ces estimateurs sont particulièrement utiles pour l'optimisation. En effet, les modèles de coût reposent généralement sur de telles estimations pour choisir quel sera le terme le plus efficace parmi plusieurs termes équivalents. Pour cette estimation, nous nous intéressons tout particulièrement au fragment conjonctif de la µ-algèbre (ce qui correspond au fragment Basic Graph Pattern de SPARQL). Notre nouvelle estimation de cardinalité s'appuie sur des statistiques sur

[11] Spark est un framework open source de calcul distribué. Il s'agit d'un ensemble d'outils et de composants logiciels structurés selon une architecture définie. Développé à l'université de Californie à Berkeley, Spark est un projet de la fondation Apache. C'est un cadre applicatif de traitements des mégadonnées (big data) pour effectuer des analyses complexes à grande échelle.

les données et a été implémentée dans SPARQLGX. Nos expériences montrent que cette méthode permet de grandement accélérer l'évaluation de SPARQL sur SPARQLGX et d'être plus robuste aux pannes des nœuds de calcul.

Notes et références

Raisonner sur des données en présence d'ontologies

Ouvrage :

Serge Abiteboul, Ioana Manolescu, Philippe Rigaux, Marie-Christine Rousset, Pierre Senellart, Web Data Management, Cambridge University Press, Cambridge (UK), 2012

Thèses :

Shadi Baghernezhad Tabasi, Interactive ontology modeling and updating: application to simulation-based training in Medicine, Thèse Université Grenoble Alpes, 2021

Adam Sanchez, Large-scale ontology-based data analytics: application to the SIDES 3.0 training platform in Medicine, Thèse Université Grenoble Alpes, 2023

Articles :

Olivier Palombi, Federico Ulliana, Valentin Favier, Jean-Claude Léon and Marie-Christine Rousset, My Corporis Fabrica: an ontology-based tool for reasoning and querying on complex anatomical models, *Journal of Biomedical Semantics* 5:20, 2014

Pierre-Yves Rabattu, Benoit Massé, Federico Ulliana, Marie-Christine Rousset, Damien Rohmer, Jean-Claude Léon and Olivier Palombi, My Corporis Fabrica Embryo: An ontology-based 3D spatio-temporal modeling of human embryo development, *Journal of Biomedical Semantics* 6:36, 2015

Olivier Palombi, Fabrice Jouanot, Nafissetou Nziengam, Behrooz Omidvar-Tehrani, Marie-Christine Rousset, Adam Sanchez, OntoSIDES: Ontology-based student progress monitoring on the national evaluation system of French Medical Schools, *Artificial Intelligence in Medicine 96*, 2019

Hétérogénéité sémantique et alignement d'ontologies

Ouvrages :

Isabel Cruz, Stefan Decker, Jérôme Euzenat, Deborah McGuinness (eds), The emerging semantic web, IOS press, Amsterdam (NL), 302p., 2002

Jérôme Euzenat (ed), Research challenges and perspectives of the Semantic web, EU-NSF Strategic report, ERCIM, Sophia Antipolis (FR), 82p., 2002

Jérôme Euzenat, Pavel Shvaiko, Ontology matching, Springer-Verlag, Heidelberg (DE), (2nd edition 520p., 2013), 2007

Thèses :

Zhengjie Fan, Concise pattern learning for RDF data sets interlinking, Thèse d'informatique, Université de Grenoble, Grenoble (FR), 2014

Armen Inants, Qualitative calculi with heterogeneous universes, Thèse d'informatique, Université de Grenoble, Grenoble (FR), 2016

Khadija Jradeh, Optimised tableau algorithms for reasoning in the description logic ALC extended with link keys, Thèse d'informatique, Université de Grenoble, Grenoble (FR), 2022

Sébastien Laborie, Adaptation sémantique de documents multimédia, Thèse d'informatique, Université Joseph Fourier, Grenoble (FR), 2008

Tatiana Lesnikova, RDF data interlinking: evaluation of cross-lingual methods, Thèse d'informatique, Université de Grenoble, Grenoble (FR), 2016

Antoine Zimmermann, Sémantique des réseaux de connaissances: gestion de l'hétérogénéité fondée sur le principe de médiation, Thèse d'informatique, Université Joseph Fourier, Grenoble (FR), 2008

Articles :

Nacira Abbas, Jérôme David, Amedeo Napoli, Discovery of link keys in RDF data based on pattern structures: preliminary steps, in: Proc. 15th International conference on Concept Lattices and their Applications (CLA), Tallinn (EE), pp235-246, 2020

Mustafa Al-Bakri, Manuel Atencia, Steffen Lalande, Marie-Christine Rousset, Inferring same-as facts from linked data: an iterative import-by-query approach, Proc. 29th conference on Conference on Artificial Intelligence (AAAI), Austin (TX US), pp9-15, 2015

Mustafa Al-Bakri, Manuel Atencia, Jérôme David, Steffen Lalande, Marie-Christine Rousset, Uncertainty-sensitive reasoning for inferring sameAs facts in linked data, Proc. 22nd european conference on artificial intelligence (ECAI), Der Haague (NL), pp698-706, 2016

Manuel Atencia, Alexander Borgida, Jérôme Euzenat, Chiara Ghidini, Luciano Serafini, A formal semantics for weighted ontology mappings, in: Proc. 11th conference on International semantic web conference (ISWC), Boston (MA US), *Lecture notes in computer science* 7649:17-33, 2012

Manuel Atencia, Jérôme David, François Scharffe, Keys and pseudo-keys detection for web datasets cleansing and interlinking, in: Proc. 18th international conference on knowledge engineering and knowledge management (EKAW), Galway (IE), *Lecture notes in computer science* 7603:144-153, 2012

Manuel Atencia, Marco Schorlemmer, An interaction-based approach to semantic alignment, *Journal of web semantics* 13:131-147, 2012

Manuel Atencia, Michel Chein, Madalina Croitoru, Jérôme David, Michel Leclère, Nathalie Pernelle, Fatiha Saïs, François Scharffe, Danai Symeonidou, Defining key semantics for the RDF datasets: experiments and evaluations, in: Proc. 21st International Conference on Conceptual Structures (ICCS), Iasi (RO), *Lecture notes in artificial intelligence* 8577:65-78, 2014

Manuel Atencia, Jérôme David, Jérôme Euzenat, Data interlinking through robust linkkey extraction, in: Proc. 21st European conference on artificial intelligence (ECAI), Praha (CZ), pp15-20, 2014

Manuel Atencia, Jérôme David, Jérôme Euzenat, Several link keys are better than one, or extracting disjunctions of link key candidates, in: Proc. 10th ACM international conference on knowledge capture (K-Cap), Marina del Rey (CA US), pp61-68, 2019

Manuel Atencia, Jérôme David, Jérôme Euzenat, Amedeo Napoli, Jérémy Vizzini, Link key candidate extraction with relational concept analysis, *Discrete applied mathematics* 273:2-20, 2020

Manuel Atencia, Jérôme David, Jérôme Euzenat, On the relation between keys and link keys for data interlinking, *Semantic web journal* 12(4):547-567, 2021

Jérôme David, Jérôme Euzenat, François Scharffe, Cássia Trojahn dos Santos, The Alignment API 4.0, *Semantic web journal* 2(1):3-10, 2011

Jérôme Euzenat, Petko Valtchev, Similarity-based ontology alignment in OWL-Lite, in: Ramon López de Mantaras, Lorenza Saitta (eds), Proc. 16th European conference on artificial intelligence (ECAI), Valencia (ES), pp333-337, 2004

Jérôme Euzenat, Semantic precision and recall for ontology alignment evaluation, in: Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI), Hyderabad (IN), pp348-353, 2007

Jérôme Euzenat, Christian Meilicke, Pavel Shvaiko, Heiner Stuckenschmidt, Cássia Trojahn dos Santos, Ontology Alignment Evaluation Initiative: six years of experience, *Journal on data semantics* XV(6720):158-192, 2011

Jérôme Euzenat, Maria Rosoiu, Cássia Trojahn dos Santos, Ontology matching benchmarks: generation, stability, and discriminability, *Journal of web semantics* 21:30-48, 2013

Jérôme Euzenat, Revision in networks of ontologies, *Artificial intelligence* 228:195-216, 2015

Angela Locoro, Jérôme David, Jérôme Euzenat, Context-based matching: design of a flexible framework and experiment, *Journal on data semantics* 3(1):25-46, 2014

Pavel Shvaiko, Jérôme Euzenat, Ontology matching: state of the art and future challenges, *IEEE Transactions on knowledge and data engineering* 25(1):158-176, 2013

Antoine Zimmermann, Jérôme Euzenat, Three semantics for distributed systems and their relations with alignment composition, in: Proc. 5th International semantic web conference (ISWC), Athens (GA US), *Lecture notes in computer science* 4273, 2006), pp16-29, 2006

Extension et optimisation de requêtes SPARQL

Thèses :

Faisal Alkhateeb, Querying RDF(S) with regular expressions, Thèse d'informatique, Université Joseph Fourier, Grenoble (FR), June 2008

Melisachew Wudage Chekol: Static Analysis of Semantic Web Queries. (Analyse Statique de Requête pour le Web Sémantique). Grenoble Alpes University, France, 2012.

Sarah Chlyah, Fondements algébriques pour l'optimisation de la programmation itérative avec des collections de données distribuées. Grenoble Alpes University, France, 2022.

Damien Graux, On the Efficient Distributed Evaluation of SPARQL Queries. (De l'évaluation répartie et efficace de requêtes SPARQL). Grenoble Alpes University, France, 2016.

Louis Jachiet, On the foundations for the compilation of web data queries: optimization and distributed evaluation of SPARQL. (Sur la compilation des langages de requêtes pour le web des données : optimisation et évaluation distribuée de SPARQL). Grenoble Alpes University, France, 2018.

Muideen Lawal, On Cost Estimation for the Recursive Relational Algebra. (Sur l'estimation des coûts pour l'algèbre relationnelle récursive). Grenoble Alpes University, France, 2021.

Articles :

Faisal Alkhateeb, Jean-François Baget, Jérôme Euzenat, Extending SPARQL with regular expression patterns (for querying RDF), *Journal of web semantics* 7(2):57-73, 2009

Faisal Alkhateeb, Jérôme Euzenat, Querying RDF data, in: SHERIF SAKR, ERIC PARDEDE (eds), Graph data management: techniques and applications, IGI Global, Hershey (PA US), pp337-356, 2012

Melisachew Wudage Chekol, Jérôme Euzenat, Pierre Genevès, Nabil Layaïda, SPARQL Query Containment Under Schema. *Journal on data semantics* 7(3): 133-154, 2018

Louis Jachiet, Pierre Genevès, Nils Gesbert, Nabil Layaïda, On the Optimization of Recursive Relational Queries: Application to Graph Queries. SIGMOD Conference, pp681-697, 2020