Fédération multi-sources en neurosciences : intégration de données relationnelles et sémantiques

Alban Gaignard, Johan Montagnat, Catherine Faron Zucker, Olivier Corby

Fédération multi-sources en neurosciences : intégration de données relationnelles et sémantiques

Alban Gaignard[1], Johan Montagnat[1],
Catherine Faron Zucker[1], Olivier Corby[2]

[1] CNRS / UNS, laboratoire I3S, équipe MODALIS, Sophia Antipolis, France
{alban.gaignard, johan, faron}@i3s.unice.fr

[2] INRIA Sophia Antipolis, équipe Wimmics, France
olivier.corby@inria.fr

Résumé : La fédération et l'interrogation multi-sources de données est un besoin croissant. En neurosciences collaboratives, les entrepôts de données sont hétérogènes et ne peuvent être relocalisés hors des sites d'origine, pour des raisons historiques, juridiques ou éthiques. Cet article présente un système de recherche d'informations qui s'interface à des entrepôts de données multiples, hétérogènes et distribués. Ce système est évalué dans le cadre d'une plateforme de neurosciences collaboratives dédiée aux études cliniques multi-centriques en termes d'utilisabilité et de performance.
Mots-clés : Web de données, Entrepôts fédérés, Intégration de données.

1 Introduction

Avec le développement de l'Internet et de ses usages, de nouvelles sources de données apparaissent tous les jours. Les données publiées en ligne adoptent en général les standards du Web pour la représentation des connaissances, mais peuvent également être contraintes par des formats de représentation plus anciens tels que les formats relationnels. Le concept de *Linked Data* émerge du besoin de fédérer de telles sources de données à l'échelle du Web et pose de nouveaux défis liés à l'interrogation transparente de sources de données multiples et distribuées, dans un référentiel sémantique unifié, en particulier lorsque les entrepôts héritent de modèles hétérogènes.

Une approche souvent envisagée est la centralisation des données (*data warehousing*). Elle consiste à importer les données issues des sources distribuées dans un entrepôt unique. Bien que simple, cette solution a de nombreux inconvénients tels que la nécessité de transformer les données originelles, la périodicité des actualisations de l'entrepôt, son passage à l'échelle ou sa tolérance aux pannes. Dans certains domaines d'application tels que les neurosciences il n'est pas possible de créer de tels entrepôts centralisés pour des raisons juridiques ou éthiques. Une solution alternative consiste à

proposer un mécanisme de fédération dynamique de données distribuées qui aligne des modèles de données hétérogènes pour fournir une vue unifiée sur ces données. Dans cette approche, les requêtes sont propagées sur les sources de données et les résultats sont intégrés côté utilisateur. Cela pose des problèmes liés à l'adaptation des requêtes à chacune des sources, et à la transformation à la volée de données tout en préservant leur sémantique et en limitant le coût des communications.

Dans cet article, nous adoptons une approche de fédération multi-sources dynamique répondant aux besoins des neuro-scientifiques. Cette communauté, généralement impliquée dans des études multi-centriques, exploite des données fragmentées dans différents centres de recherche sous la forme d'entrepôts pré-existants et hétérogènes (Gibaud *et al.*, 2011). Nous avons développé, dans le cadre de KGRAM (Corby & Zucker, 2010), un système basé sur les modèles et les techniques du Web Sémantique permettant la fédération transparente de données hétérogènes et multi-sources. KGRAM est une plateforme générique et ne nécessite pas de connaissance a priori sur le contenu des sources de données. Il est robuste, car il s'adapte aux changements de topologie liés à la disponibilité variable des sources.

La section 2 introduit KGRAM et ses possibilités d'optimisation. La section 3 décrit une application dans le domaine des neurosciences, ainsi que des résultats expérimentaux. Finalement, la section 4 discute des travaux comparables et propose quelques perspectives.

2 KGRAM pour la distribution de requêtes sémantiques

KGRAM–*Knowledge Graph Abstract Machine*– est un environnement intégré à Corese [1], qui vise à représenter, interroger et raisonner sur des graphes de connaissances. KGRAM est conçu comme l'interprète d'un langage abstrait qui généralise SPARQL 1.1 (agrégats, requêtes imbriquées, négations, chemins) et permet l'interrogation de sources de données hétérogènes (comme RDF, XML, SQL), à la condition que ces sources publient une vue des données sous la forme d'un graphe. Dans KGRAM, l'évaluation de requêtes SPARQL consiste à chercher des nœuds et des arcs dans un graphe de connaissances publié par un *producteur*. Pour interroger des données interconnectées (*Linked Data*), KGRAM introduit la notion de *méta-producteur* rendant possible l'énumération de nœuds et d'arcs provenant de plusieurs *producteurs*.

Pour interroger des entrepôts distribués, nous avons étendu KGRAM avec un service web permettant d'invoquer un *producteur* distant. A travers ce service, un *méta-producteur* interroge plusieurs sources distantes de manière asynchrone puis fusionne les résultats obtenus dans un graphe de connaissances résultat. L'algorithme d'interrogation distribuée consiste, pour chacun des arcs requête formant la requête SPARQL initiale, à interroger en parallèle chacun des *producteurs* distants. Le *méta-producteur* attend la terminaison de chacun des *producteurs* distants au travers d'une barrière de synchronisation et accumule les résultats obtenus. L'arc requête successeur peut alors être traité avec le même procédé. Pour assouplir la contrainte de synchronisation nous

1. http ://wimmics.inria.fr/corese

avons adopté une approche de type *pipeline* avec une file d'attente synchronisée qui permet de traiter *a posteriori* les résultats dès qu'ils sont disponibles.

Pour améliorer les performances de l'interrogation distribuée, nous avons intégré des optimisations statiques et dynamiques permettant de réduire drastiquement le coût des communications entre le *méta-producteur* et les *producteurs* distants. Par exemple, les contraintes sur des valeurs de variables sont transmises aux *producteurs* par la génération de clauses *Filter* applicables aux sous-requêtes SPARQL qui permettent de filtrer les résultats non pertinents directement à leur source. Nous avons également intégré une optimisation dynamique permettant d'exploiter les résultats intermédiaires dans les sous-requêtes envoyées aux *producteurs* distants (*bind joins*). Cela réduit le coût de traitement des requêtes car il y a moins de triplets candidats et, par là même, cela réduit le coût des communications.

Finalement, l'extensibilité de KGRAM nous a permis de développer un médiateur capable d'énumérer le contenu de bases de données relationnelles sous la forme de triplets directement intégrables dans le graphe de connaissances résultat. Ce *producteur-médiateur* est en mesure de générer, à la volée, une requête SQL, et d'associer aux variables de la requête SPARQL, les résultats provenant d'une source relationnelle.

3 Applications en neurosciences

3.1 La plate-forme NeuroLOG

NeuroLOG (Gibaud *et al.*, 2011) est un intergiciel visant à fédérer des données et des services de traitement de données de 5 centres de recherche en France qui collaborent dans le domaine des neurosciences. L'objectif de la plateforme est de s'adapter de manière non invasive aux environnements pré-existants de chacun des centres pour que chacun puisse conserver son autonomie sur la gestion des ressources hébergées, tout en bénéficiant des possibilités de collaboration (études multi-centriques). Chaque centre expose des données image brutes, ainsi que leurs méta-données associées (description du contenu des données, des protocoles d'acquisition, des tests neuro-psychologiques associés, etc.). Ces méta-données sont en général gérées dans des bases relationnelles. Au vu des contraintes d'autonomie, du volume et de la sensibilité des données, les approches centralisées sont inappropriées. Nous avons développé dans NeuroLOG une fédération de données relationnelles dirigée par une approche ontologique. De plus, afin d'exploiter la richesse de l'ontologie et l'expressivité du langage SPARQL, nous avons complété la plateforme avec le moteur d'interrogation distribué de KGRAM.

La figure 1 illustre l'architecture de l'intergiciel NeuroLOG (en violet) complétée par le moteur de recherche d'information distribué KGRAM (en vert). Chacun des sites collaborant dans la plateforme, gère une base de données indépendante. L'outil commercial DataFederator est utilisé pour fédérer dynamiquement les bases relationnelles de chacun des sites au travers d'une vue relationnelle unifiée. Cet outil inclut une couche de médiation dans laquelle les schémas hétérogènes, provenant des sites collaborant dans la plateforme, sont alignés sur le schéma fédéré, qui dérive de l'ontologie de domaine OntoNeuroLOG (également développée dans le cadre du projet NeuroLOG). Un producteur KGRAM a été déployé sur chacun des sites de la fédération et expose, le contenu

des données au format RDF. En fonction de la configuration du site, le producteur s'interface soit directement à la base de données relationnelle (option ①), soit à un entrepôt RDF créé périodiquement à partir de l'outil MetaMORPHOSES (option ②).

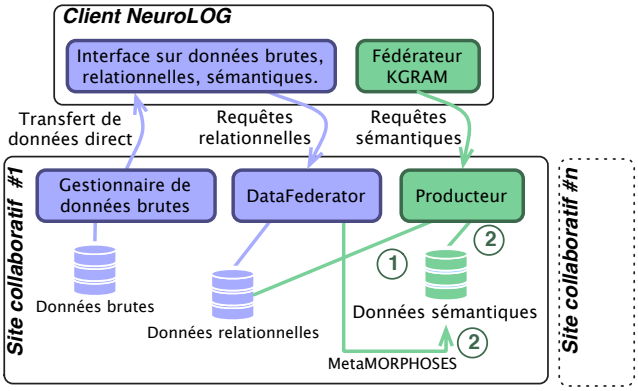

FIGURE 1 – Médiation de données brutes, relationnelles et sémantiques avec l'intergiciel NeuroLOG.

3.2 Résultats expérimentaux

Nous évaluons brièvement, dans cette section, la pertinence d'une fédération sémantique par rapport à une fédération purement relationnelle. Sur un plan qualitatif, les requêtes exprimées en SPARQL bénéficient tout d'abord des possibilités d'inférences offertes par KGRAM telles que la subsomption (*RDFS entailment*), la transitivité, l'inversion ou la symétrie de propriétés. Ces inférences ne sont pas directement réalisables au travers de modèles purement relationnels. Cependant, lorsque des bases relationnelles participent à une fédération sémantique telle qu'envisagée dans cet article, elles peuvent alors bénéficier de telles possibilités d'inférence. Par ailleurs, la conception de requêtes est plus intuitive avec les langages à base de graphes (tels que SPARQL 1.1 incluant les expressions de chemins) qu'avec des langage relationnels tels que SQL. En effet, la navigation au travers des relations entre données est explicite avec SPARQL alors qu'elle est implicite avec SQL et nécessite en général des jointures intermédiaires (tables de jointure).

La requête présentée dans la figure 2 (sous forme SPARQL et SQL) illustre un besoin clinique réel : la recherche d'information dans des bases inter-connectées de connaissances en neuro-imagerie, la requête SPARQL exploitant directement les concepts et propriétés d'OntoNeuroLOG, et la requête relationnelle s'appuyant sur le schéma directement dérivé d'OntoNeuroLOG. En particulier, l'objectif de cette requête est de chercher des images (*?datasetName*), acquises avec du Gadolinium comme agent de contraste, et leurs patients (*?patient*) associés (jointure ligne 4 de la requête SPARQL) dans le contexte d'une étude multi-centrique (*?study*) sur la sclérose en plaques.

```
1 SELECT distinct ?patient ?study ?datasetName WHERE {
2   ?patient iec:is−referred−by/exp:has−for−name ?datasetName .
3   ?patient subject:has−for−identifier ?clinID .
4   ?study study:involves−as−patient ?patient .
5   FILTER (regex(?clinID, 'MS') && regex(?datasetName, 'GADO')) }
```

```
1 SELECT Subject.subject_id, Subject.subject_common_identifier, Dataset.name
2   FROM Study, Subject, Dataset, Rel_Subject_Study WHERE
3     Rel_Subject_Study.Subject_subject_id = Subject.subject_id AND
4         Rel_Subject_Study.Study_study_id = Study.study_id AND
5         Dataset.Subject_subject_id = Subject.subject_id AND
6         Subject.subject_common_identifier LIKE '%MS%' AND
7         Dataset.name LIKE '%GADO%'
```

FIGURE 2 – Comparaison des langages SPARQL 1.1 (en haut) incluant une expression de chemin (ligne 2) et SQL (en bas) pour une même recherche d'informations en neuro-imagerie.

La figure 2 montre que les requêtes basées sur la navigation de graphe ne sont pas facilement exprimables en SQL. Elles peuvent en effet nécessiter des tables de jointures (table *Rel_Subject_Study*) et sont plus difficiles à exprimer avec des données relationnelles. KGRAM est un interprète SPARQL 1.1 qui facilite la recherche dans les fédérations sémantiques, en incluant éventuellement des sources relationnelles.

En utilisant la plateforme NeuroLOG, nous avons comparé la performance d'une fédération purement relationnelle (médiateur DataFederator de SAP) avec une fédération sémantique (KGRAM, sur des bases RDF). Nous avons observé que pour des requêtes distribuées très sélectives (peu de jointures, 5 résultats) nous obtenons des résultats moyennés légèrement meilleurs pour la fédération sémantique (0.6s contre 1.5s en relationnel). Pour des requêtes plus coûteuses en termes de jointures (336 sous-requêtes SPARQL) les résultats moyennés sont meilleurs pour la fédération relationnelle (3s contre 6s en sémantique) mais ceux de la fédération sémantique conservent cependant un même ordre de grandeur, et restent acceptables.

4 Discussion et conclusion

Le W3C a proposé récemment une extension [2] du language SPARQL visant à interroger des sources distribuées de données RDF, mais cette approche nécessite de modifier la requête pour indiquer explicitement les sources de données adaptées à une sous-partie de la requête. Cette approche n'est pas adaptée dans pour des fédérations dynamiques où la disponibilité des sites n'est pas garantie. La fédération transparente de données sémantiques est abordée par DARQ (Quilitz & Leser, 2008), SPLENDID (Görlitz & Staab, 2011) ou FedX (Schwarte *et al.*, 2011) sous l'angle de la performance avec un ensemble d'optimisations statiques et dynamiques. KGRAM intègre déjà certaines optimisations réduisant le coût de l'évaluation distribuée mais pourrait bénéficier d'autres

2. http ://www.w3.org/TR/sparql11-federated-query

améliorations (groupage de sous-requêtes, index sur le contenu et la structure des bases, etc). Par ailleurs, le médiateur SQL proposé dans ce travail pourrait bénéficier de l'expérience du projet D2RQ (Bizer & Cyganiak, 2007) et des avancées récentes en terme de standardisation (R2RML [3]). Mais à notre connaissance, aucune des approches récentes permettant la fédération multi-source transparente de données sémantiques n'aborde à la fois les problématiques d'hétérogénéité et de performance, ce qui rend l'approche générique et extensible envisagée avec KGRAM très encourageante.

En résumé, la fédération cohérente de données multi-sources est un besoin grandissant dans le domaine des neurosciences, pour développer des études multi-centriques à grande échelle. Les modèles du Web Sémantique permettent de raisonner sur des données distribuées et inter-connectées mais des efforts restent à accomplir pour permettre l'alignement sémantique et l'interrogation de sources multiples et hétérogènes de manière transparente. KGRAM facilite la mise en place de telles fédérations en répondant aux problématiques de performance et d'hétérogénéité. Au delà des aspects de performance, l'environnement d'interrogation bénéficie de l'expressivité de SPARQL 1.1, des possibilités d'inférence associées, et potentiellement des connaissances capitalisées dans les ontologies de domaine. Bien qu'appliqué aux neurosciences collaboratives, ce travail est suffisamment général pour trouver de nombreuses applications en e-Santé / e-Sciences, où les systèmes distribués, l'ingénierie des connaissances, et les ontologies de domaines prennent chaque jour une place plus importante.

Remerciements

Ce travail a été partiellement financé par l'Agence Nationale de la Recherche (projet NeuroLOG ANR-06-TLOG-024).

Références

BIZER C. & CYGANIAK R. (2007). D2RQ - Lessons Learned. *W3C Workshop on RDF Access to Relational Databases*.

CORBY O. & ZUCKER C. F. (2010). The kgram abstract machine for knowledge graph querying. *Web Intelligence and Intelligent Agent Technology, IEEE/WIC/ACM International Conference on*, **1**, 338–341.

GIBAUD B., AHMAD F., BARILLOT C., MICHEL F., WALI B., BATRANCOURT B., DOJAT M., GIRARD P., GAIGNARD A., LINGRAND D., MONTAGNAT J., ROJAS BALDERRAMA J., MALANDAIN G., PENNEC X., GODARD D., KASSEL G. & PÉLÉGRINI-ISSAC M. (2011). A federated system for sharing and reuse of images and image processing tools in neuroimaging. In *Computer Assisted Radiology and Surgery (CARS'11)*, Berlin, Germany.

GÖRLITZ O. & STAAB S. (2011). SPLENDID : SPARQL Endpoint Federation Exploiting VOID Descriptions. In *Proceedings of the 2nd International Workshop on Consuming Linked Data*, Bonn, Germany.

QUILITZ B. & LESER U. (2008). Querying distributed rdf data sources with sparql. *The Semantic Web Research and Applications*, **5021**, 524–538.

SCHWARTE A., HAASE P., HOSE K., SCHENKEL R. & SCHMIDT M. (2011). Fedx : optimization techniques for federated query processing on linked data. In *Proceedings of the 10th international conference on The semantic web - Volume Part I*, ISWC'11, p. 601–616, Berlin, Heidelberg : Springer-Verlag.

3. http ://www.w3.org/TR/r2rml