Mission pour l'Interdisciplinarité du CNRS MASTODONS

Défi Grandes Masses de Données Scientifiques

CrEDIBLE

ConnaissancEs Distribuées en Imagerie BiomédicaLE

http://credible.i3s.unice.fr

Document ID. CrEDIBLE-12-2-v1

Novembre 2012

Bilan de l'année 2012

C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, J. Montagnat

Résumé

Ce document décrit les objectifs du projet CrEDIBLE et le travail réalisé au cours de l'année 2012. Il rappelle la vision scientifique du projet et ses principales orientations. Il résume l'atelier multidisciplinaire organisé à Sophia Antipolis au mois d'octobre 2012 sur les thématiques du projet et donne un aperçu des études bibliographiques menées. La conclusion donne les orientations du travail envisagé pour les prochaines années.

Table des matières

1 Participants au projet

- CNRS, laboratoire I3S (UMR7271), équipe MODALIS
- INRIA/CNRS/UNS, laboratoire I3S (UMR 7271), équipe Wimmics
- INSERM U1099, laboratoire LTSI, équipe MediCIS
- U. Picardie, laboratoire MIS, équipe Connaissances
- CNRS/INSERM/INSA/U. Lyon 1, laboratoire CREATIS (UMR 5220 / U1044)

2 Motivations

A l'instar de ce qui se passe dans de nombreux autres domaines, la communauté médicale fait face à un accroissement du volume des données acquises sous forme numérique qui va en s'accélérant. Les cohortes de données ainsi accumulées constituent un capital précieux pour répondre aux défis actuels de la médecine puisqu'elles permettent de réaliser des études sur des populations à grande échelle (statistiques ou épidémiologiques), d'assurer un meilleur suivi des patients au cours du temps ou d'étudier des cas d'occurrence rare, par exemple. Elles soulèvent cependant des défis considérables en raison de la quantité de données à manipuler, de la complexité des structures de données nécessaires à la représentation de l'information médicale et à la nécessité de disposer de référentiels et de métadonnées explicitant sa sémantique.

En outre, la distribution des données médicales est à la fois un état de fait, en raison de la multiplication des instruments d'acquisition des données numériques publiant leurs données en ligne systématiquement, et nécessaire, au vu des quantités de données exploitées. Le besoin de distribution est renforcé par les contraintes juridiques et éthiques s'appliquant à ce type de données. La distribution des calculs s'ensuit naturellement, au vu des quantités de données à analyser, ou simplement pour faire face aux difficultés de transfert de certaines données sensibles.

Actuellement, on assiste au déploiement de nombreux entrepôts d'images médicales dédiés à la recherche clinique et translationnelle dans des centres disposant de ressources multiples pour l'acquisition, le stockage et l'analyse de ces données. La création de référentiels numériques médicaux, la représentation de données médicales de différentes natures à l'aide de ces référentiels, l'interrogation conjointe d'entrepôts de données distants et le traitement distribué de cohortes de données sont donc des éléments clés du développement du secteur de la recherche médicale.

La plateforme distribuée NeuroLOG (ANR-06-TLOG-024) a ainsi été développée pour faciliter la mise en œuvre d'études multicentriques en neurosciences (notamment dans les domaines de la sclérose en plaques, des tumeurs cérébrales, des accidents vasculaires cérébraux et de la maladie d'Alzheimer). Son intergiciel permet de fédérer des sources de données hétérogènes disponibles dans 5 centres de neurosciences répartis en France. Il intègre une interface avec la grille de calcul Européenne EGI. Afin d'établir un référentiel commun, une ontologie de domaine nommée OntoNeuroLOG a été développée dans ce cadre. Elle couvre les aspects d'acquisition et de représentation des images médicales, de tests neuro-psychologiques et d'études cliniques. Ce projet exploratoire a permis d'étudier la faisabilité et les conditions d'usage d'entrepôts distribués de données médicales. Le projet CrEDIBLE a pour objectif de prolonger cet effort, en étudiant l'intégration plus systématique des données médicales dans des entrepôts distribués de connaissances, la constitution de corpus d'étude par l'interrogation conjointe de tels entrepôts, leur analyse à travers des chaînes de traitement adaptées à la manipulation de grands ensembles de données, et enfin l'inférence de nouvelles données médicales.

3 Vision scientifique du projet

La gestion de données médicales distribuées telle qu'envisagée dans le projet CrEDIBLE recouvre :

- La fusion (virtuelle) d'entrepôts physiquement distribués mais devant apparaître pour leurs exploitants comme une entité unique et cohérente.
- L'alignement sémantique de sources de données hétérogènes, qui n'ont souvent pas été conçues pour être exploitées conjointement.
- La description d'ensembles de données distribuées, définis par l'intermédiaire de requêtes qui peuvent s'appliquer sur l'ensemble de la fédération.
- L'analyse ou la transformation des ensembles de données ainsi constitués par l'intermédiaire de langages pilotés par les données.

Le travail proposé intègre donc les moyens d'aligner des entrepôts de données hétérogènes (médiation) et de les unifier (fédération), notamment à travers des outils d'interrogation (requêtes distribuées), et de les analyser (flux de données) sur des infrastructures de calcul distribuées. Il aborde donc plus particulièrement les axes suivants de l'appel à manifestation d'intérêt MASTODONS:

2. Calcul sur des grands volumes, parallélisme dirigé par les données.
3. Recherche de grandes masses de données.
7. Préservation/archivage des données pour les générations futures.

Les principaux verrous scientifiques abordés sont :

1. la **représentation sémantique** des données d'imagerie médicale fondée sur des ontologies des différents domaines concernés (recherche clinique, imagerie, traitement des images, marqueurs quantitatifs associés à des structures anatomiques ou processus physiologiques ou physiopathologiques, scores cliniques…) ;
2. la gestion de **sources de données hétérogènes** par des mécanismes de médiation dynamiques fondés sur la sémantique ;
3. la **fédération d'entrepôts distribués** par le biais de mécanismes de requêtes applicables à l'ensemble de la fédération (distribution, réécriture) ;
4. la **performance** des requêtes distribuées ; et
5. la gestion de **flots de calculs distribués** sur les cohortes de données ainsi constituées, décrits par l'intermédiaire de langages pilotés par les données.

Ces thématiques principales nécessiteront également d'aborder :

6. la **cohérence des données distribuées** afin de lier les instances relatives à une même entité physique potentiellement distribuées dans plusieurs entrepôts, et d'assurer la cohérence des connaissances qui peuvent apparaître sous plusieurs représentations dans différents entrepôts; et
7. le **contrôle d'accès aux données** les plus sensibles, qui peut être envisagé comme une étape du processus dynamique de médiation.

4 Travail réalisé

Le travail réalisé au cours des premiers mois du projet a consisté en :

- une étude élargie du domaine abordé, en particulier des ressources ontologiques et des moteurs d'interrogation de bases de données distribuées ; et
- un prototype de moteur de recherche dans des d'entrepôts sémantiques distribués, basé sur le logiciel KGRAM (*Knowledge Graph Abstract Machine*) développé à l'INRIA Sophia Antipolis.

Pour assurer une bonne prise en compte des travaux existants dans ce domaine et une bonne diffusion du projet CrEDIBLE, un atelier multidisciplinaire a été organisé à

Sophia Antipolis en octobre 2012. Il faisait intervenir des spécialistes des domaines de l'intégration des données, de la modélisation sémantique, des modèles de représentation de données, du raisonnement, et des flux de calcul sémantiques.

4.1 Atelier multidisciplinaire

Le projet CrEDIBLE a organisé un atelier de travail multidisciplinaire international qui s'est déroulé entre le 15 et 17 octobre à Sophia Antipolis. Cet atelier avait pour but de réunir des spécialistes de toutes les disciplines concernées par la mise en œuvre de systèmes de gestion de données médicales distribuées hétérogènes (incluant la représentation des données, sémantique, distribution, intégration et fusion de données, information de provenance faisant le lien entre traitements appliqués et données stockées dans les bases) afin d'obtenir une vision aussi complète que possible de ce domaine complexe, d'en analyser les besoins, de parcourir les méthodes et technologies existantes et de discuter des questions scientifiques qu'il soulève.

Cet atelier a été organisé en quatre sessions thématiques qui ont été l'occasion de faire intervenir une vingtaine d'orateurs invités selon le programme suivant :

Session 1, Data integration methods and Tools

- **J. Montagnat** (CNRS) - Feedback from the NeuroLOG project
- **C. Daniel** (APHP / INSERM, Paris) - Electronic Health Records for Clinical Research
- **S. Murphy** (Massachusetts General Hospital / Harvard Medical School) - Instrumenting the Health Care Enterprise for Discovery Research
- **O. Corcho** (U. Polytechnic Madrid) - Distributed queries, data médiation
- **P. Grenon** (European Bioinformatics Institute) - Ontology based knowledge management of biomedical models and data
- **O. Corby** (INRIA, Sophia Antipolis) - KGRAM abstract machine for knowledge data management

Session 2: Ontologies, semantic modeling

- **C. Masolo** (Laboratory for Applied Ontology, Trento) - DOLCE extensions
- **G. Gkoutos** (U. Cambridge) - From Systems Genetics to Translational Medicine
- **J. Charlet** (APHP / INSERM, Paris) - Relations between Ontologies and Knowledge Structure: Two Case Study
- **P. Grenon** (European Bioinformatics Institute) - Ontology for biomedical models and data
- **B. Batrancourt** (APHP / INSERM, Paris) - Ontology reuse, from NeuroLOG to CATI

Session 3: Data representation model and reasoning

- **M.-A. Aufaure** (Ecole Centrale de Paris) - Crunch and Manage Graph Data: the survival kit
- **C. Raissi** (INRIA, Nancy) - Knowledge Discovery guided by Domain Knowledge in the Big Data Era
- **K. Todorov** (INRIA, Montpellier) - Bringing Together Heterogeneous Domain Ontologies via the Construction of a Common Fuzzy Knowledge Body
- **R. Choquet** (APHP / INSERM Paris) - DebugIT: Ontology-mediated Data Integration for real-time Antibiotics Resistance Surveillance
- **S. Ferré** (U. Rennes 1) - SEWELIS: Reconciling Expressive Querying and Exploratory Search
- **P. Molli** (U. Nantes) - Live Linked Data

Session 4: Semantic workflows

- **F. Lécué** (IBM) - Composing and optimizing services in the Semantic Web
- **P. Missier** (Newcastle University) - Workflows, experimental findings, and their provenance: towards semantically rich linked data and method sharing for collaborative science
- **A. Gaignard**, **N. Cerezo** (I3S, Sophia Antipolis) - Semantic workflows: design and provenance

Les conclusions de cet atelier ont été restituées dans le document [6].

4.2 Etude technologique

Dans sa première phase, le projet CrEDIBLE s'est intéressé en particulier à l'étude des ontologies utilisées pour la représentation des connaissances en imagerie biomédicale et celle des techniques d'interrogation de bases de connaissances distribuées.

4.2.1 Représentation de connaissances médicales

En s'appuyant à la fois sur l'expérience du projet NeuroLOG concernant la conceptualisation d'instruments de mesures utilisés en neurosciences [3] et les présentations de travaux récents faites lors de l'atelier multidisciplinaire, un cadre ontologique a été défini pour rendre compte de la sémantique des données d'observation. Un document décrivant cette ontologie est en cours de rédaction [4]. Ce cadre est organisé autour de DOLCE-CORE, une nouvelle version de l'ontologie fondatrice DOLCE développée au LOA (ISTC-CNR, Trento, Italie). Il étend DOLCE-CORE avec des modules couvrant les domaines des actions, des artefacts et des documents afin de mettre les données en relation avec d'autres entités participant de situations d'observations (ex : l'observateur, l'instrument utilisé, l'instant de l'observation) et de conférer ainsi du sens à ces données. La finalité de ce cadre ontologique est double : servir de base à des développements d'ontologies d'application dans des projets réalisant la fédération d'entrepôts de données ; servir de cadre de référence pour aligner d'autres ontologies fondatrices, par exemple BFO dans le domaine biomédical.

4.2.2 Interrogation de bases de connaissances distribuées

L'interrogation d'entrepôts sémantiques distribués est une problématique qui suscite beaucoup d'intérêts aujourd'hui. La multiplicité des sources de données et l'avènement du Web sémantique conduit à l'émergence d'un « *Web of Linked Data* » qui implique de pouvoir accéder à des données géographiquement distribuées mais néanmoins liées entre elles par une sémantique rendue explicite. En outre, l'interrogation conjointe de plusieurs sources de données, avec des capacités de jointure des résultats provenant de différents entrepôts, est un moyen de fédérer un ensemble de bases de données de manière transparente, en le faisant apparaître à l'interrogateur comme une seule base virtuelle. Les entrepôts de données biomédicales sont concernés par ces deux aspects puisqu'un nombre croissant d'études cherche à corroborer des informations de différentes natures (marqueurs biomédicaux issus de l'imagerie, tests cliniques, marqueurs biologiques…) ou à exploiter des bases de données contenant un information de même nature mais distribuées sur plusieurs centres d'acquisition (fédération de centres hospitaliers au niveau régional, création de cohortes de données de grande taille pour l'analyse statistique…).

Une interrogation d'entrepôts sémantiques distribués soulève cependant des défis importants en terme de complétude des résultats provenant des requêtes réalisées (sur des entrepôts qui ne contiennent individuellement qu'une fraction des informations nécessaires à la fourniture des réponses) et de performance (alors que de très grands volumes de données peuvent transiter entre différents entrepôts).

De nombreux travaux récents se sont donc concentrés sur la réécriture de requêtes et la génération de plans de requêtes permettant d'optimiser la performance du moteur d'interrogation distribué. Des compromis sont souvent réalisés concernant l'expressivité du langage d'interrogation accessible dans ce cadre.

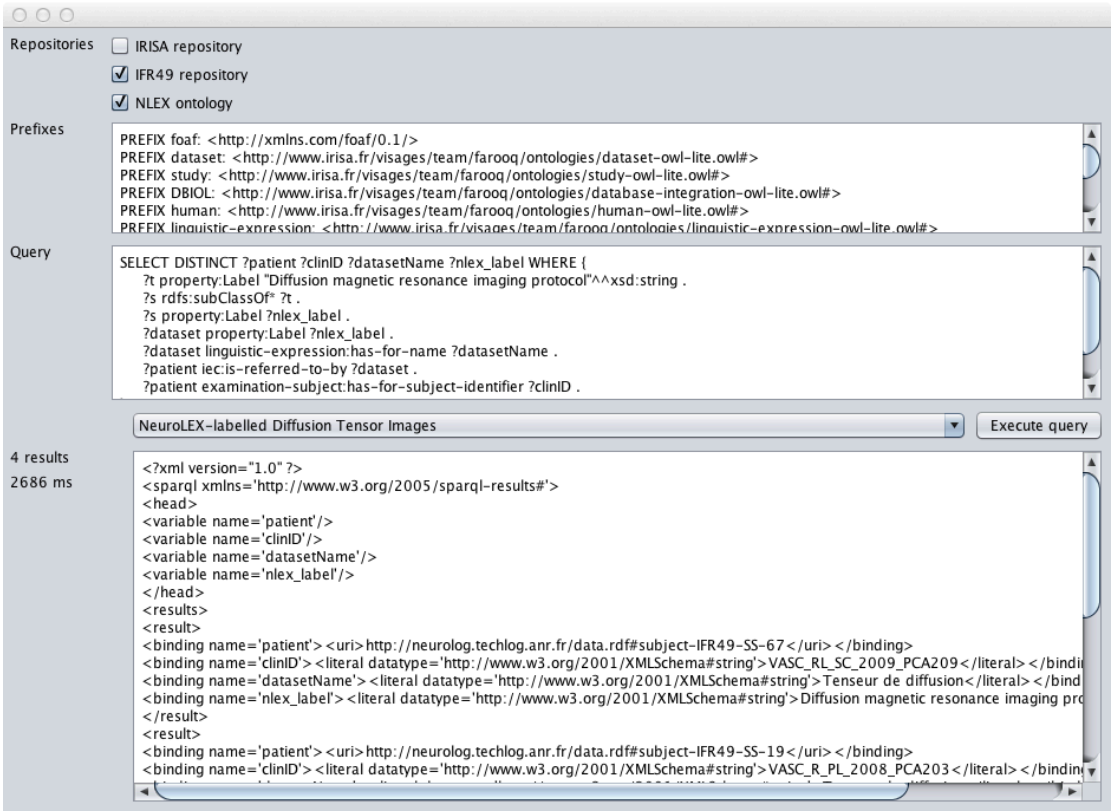
Les technologies du Web sémantique apportent cependant une grande flexibilité par la richesse du langage d'interrogation SPARQL d'une part, et la possibilité d'inférer des connaissances à travers l'application de règles déductives d'autre part. La performance des moteurs de requête distribués prime sur cette flexibilité dans de nombreux cas. Une étude bibliographique détaillée des principales approches

documentées dans la littérature a été réalisée dans le cadre du projet CrEDIBLE. Cette étude détaillée est présentée dans l'annexe de ce document.

4.3 Prototype

Un prototype de moteur de recherche dans des entrepôts sémantiques distribués est en cours de développement. Il est basé sur le logiciel KGRAM (*Knowledge Graph Abstract Machine*) développé à l'INRIA Sophia Antipolis par l'équipe Wimmics. KGRAM est un moteur d'exécution de requêtes sémantiques à l'architecture logicielle flexible qui implante le langage standard de requêtes SPARQL dans sa version 1.1. Il permet l'interrogation de sources de données qui peuvent être représentées sous forme de graphes de connaissances (RDF, mais aussi graphes conceptuels, ou d'autres sources de données transformées telles que des bases relationnelles ou XML). Le moteur KGRAM a été étendu avec des modules qui permettent l'interrogation de plusieurs entrepôts de données distants de manière concurrente [1].

Ce logiciel a été expérimenté en particulier dans le cadre de l'interrogation d'entrepôts de données pour répondre à des requêtes de recherche de données en neuro-imagerie [2]. La figure suivante montre un exemple de requête dans un prototype d'interface utilisateur.



La partie du haut permet de sélectionner les entrepôts à interroger. Dans le cadre de cet exemple, les données provenant de deux sites contribuant à la plateforme NeuroLOG (IRISA à Rennes et IFR49 à Paris) et un lexique de référence dans le domaine (NeuroLex) ont été considérés.

La partie centrale permet d'exprimer une requête dans le langage SPARQL (v1.1). Il s'agit dans cet exemple de rechercher tous les jeux de données provenant des bases de la fédération NeuroLOG, labellisés « Diffusion magnetic resonance imaging protocol » (IRM de diffusion) selon la terminologie de NeuroLex.

Enfin, la partie basse affiche les résultats obtenus, ainsi que le temps mis pour exécuter la requête.

Cet exemple démontre le bon fonctionnement du prototype et sa capacité à traiter des données provenant de bases de données réelles de structure complexe dans le domaine des neurosciences. Un travail d'approfondissement est en cours pour optimiser la performance du moteur d'exécution. L'interface avec des sources de données relationnelles et l'ajout de médiateurs scientifiques sont considérés.

5 Conclusions et perspectives

Fédérer les données médicales réparties est un besoin pour la recherche translationnelle comme pour la pratique clinique qui découle de la nature distribuée de l'infrastructure d'acquisition de ces données et du défi soulevé par les volumes de données à analyser. Une telle fédération à grande échelle ne prend de sens que si la sémantique des données entreposées est décrite précisément pour rendre possible l'alignement de différents référentiels. En outre, il est important de lier les infrastructures de calcul aux entrepôts de données biomédicales pour permettre l'analyse de celles-ci à grande échelle et donner aux scientifiques les moyens d'aborder les défis de santé actuels.

Les travaux réalisés lors des premiers mois du projet CrEDIBLE ont conduit à une analyse approfondie du domaine et des approches adoptées par ses protagonistes, notamment à travers l'atelier pluridisciplinaire organisé en octobre. Cet atelier répondait à la fois à des attentes réelles d'utilisateurs à la recherche de solutions pratiques et à une analyse des besoins de la communauté biomédicale d'un point de vu des concepteurs de méthodologies ou des développeurs de technologies. Fort de l'intérêt qu'il a suscité, une nouvelle session devrait être organisée courant 2013.

Il ressort que le principe de l'utilisation de modèles sémantiques pour la représentation des connaissances médicales est largement reconnu. Les projets de recherche clinique opérationnelle ont aujourd'hui mis en place des cohortes de données de taille très significative (jusqu'à plusieurs milliers de patients et plusieurs millions d'événements médicaux), dont l'interrogation efficace constitue un défi. Si la centralisation est encore largement considérée comme moyen pragmatique de mise en œuvre des entrepôts et de leurs interfaces de recherche, le besoin de fédérer plusieurs ressources est un besoin exprimé de manière unanime. L'interrogation distribuée des bases est souvent réalisée de manière ad hoc, peu d'outils à large diffusion existant pour aborder ce problème. L'exploitation et la réutilisation de ces données dans un cadre interdisciplinaire reste en revanche encore tout à fait marginales, et les ressources notamment ontologiques pour assurer ce type de partage (i.e. ontologies fondatrices, ontologies de domaine, ressources d'alignement) restent insuffisantes.

Le projet CrEDIBLE s'appuie sur une représentation sémantique des connaissances médicales (ontologies) et un moteur distribué d'interrogation et d'inférence sur des bases de connaissances qui implémente le standard SPARQL. Il permet de fédérer des bases de données distantes et hétérogènes via l'interrogation dynamique et la fusion de résultats provenant des différents entrepôts. Ce moteur s'adapte directement à l'interrogation de données RDF. Des travaux sont en cours pour développer une interface de médiation avec d'autres sources de données (notamment relationnelles).

A plus long terme, le projet CrEDIBLE étendra ou adaptera les ontologies de domaine nécessaires à la représentation des données considérées à travers des cas d'utilisation concrets, notamment les principales bases de données médicales mises en ligne sur le Web aujourd'hui. Le travail sur le moteur d'exécution de requêtes sera complété avec une analyse de performances et la mise en œuvre de stratégies d'optimisation.

6 Publications

[1] O. Corby, A. Gaignard, C. Faron-Zucker, J. Montagnat. "KGRAM Versatile Inference and Query Engine for the Web of Linked Data" in Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'12), Macao, China, December 2012.

[2] A. Gaignard, J. Montagnat, C. Faron-Zucker, O. Corby. "Semantic Federation of Distributed Neurodata" in Proceedings of the MICCAI Workshop on Data- and Compute-Intensive Clinical and Translational Imaging Applications (DCICTIA-MICCAI 2012), pages 41-50, Nice, France, October 2012.

[3] B. Batrancourt, M. Dojat, B. Gibaud, G. Kassel. "A multi-layer ontology of instruments for neurological, behavioral and cognitive assessments" (en cours de soumission)

[4] B. Gibaud, G. Kassel. « Sémantique des données d'observation : une approche ontologique », rapport interne CrEDIBLE-13-1-v1, à paraître.

[5] B. Gibaud, « Toward ontology-based federated systems for sharing medical images: lessons from the NeuroLOG experience ». iDash Image informatics Workshop, University of California, San Diego, La Jolla (USA), September 29, 2012. (invited talk).

[6] O. Corby, C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, J. Montagnat, « CrEDIBLE Multi-disciplinary workshop », rapport interne CrEDIBLE-12-1-v1, novembre 2012.

7 Annexe : étude bibliographique des outils d'interrogation d'entrepôts sémantiques distribués

DARQ [Quilitz and Leser, 2008] tackles semantic data integration through a federated approach. It aims at transparently federating multiple distributed and autonomous RDF data providers. The approach is based on service descriptions to allow accurate data sources selection, and on a query optimization algorithm to reduce the cost of sub-query distribution. The general process consists in (i) parsing the initial SPARQL query, then (ii) based on the service descriptions the initial SPARQL query is decomposed into sub-queries (query planning), (iii) the resulting query plan (sequence of sub-queries) is optimized and finally (iv) distributed to all data providers (query execution).

Based on service descriptions, The DARQ query planner searches for relevant sources contributing to the query results and then builds a set of suitable sub-queries. Service descriptions are represented in RDF and encompass:

(i) Data description through predicate capabilities in the form of existing predicates and their associated value constraints on subjects or objects;

(ii) Access pattern limitations [Florescu *et al.*,1999] to represent, from the data source point of view, triple patterns that must be included into a query in order to provide results;

(iii) Statistical information describing the total number of triples N, and optional information describing the capability of the source for a given predicate. This optional information covers the number of triples for a given predicate, and two triple pat- tern selectivities, the first one considering that the subject is bound, the other one considering that the object is bound.

Once available, these service descriptions are used by the query planner first, to perform source selection and second, to build sub-queries. For each basic graph pattern (BGP) constituting the initial SPARQL query, the source selection consists in matching all triple patterns against the predicate-based capabilities of the sources. These predicate-based capabilities constitute a limitation of DARQ since it is not possible to select a source if a predicate is unbound in the basic graph pattern. For example, queries containing triples patterns like ("aPerson", ?p, "aLocation") with ?p a variable predicate, cannot be processed through DARQ.

Once sources have been selected, the query planner decomposes the initial basic graph patterns into sub-queries with the following principle. A set of triple patterns is

associated to each data source. If a triple pattern matches exactly one data source, it is added to the set associated to the data source, and this set is finally processed to build a single sub-query to be sent to this precise data source. On the contrary, if a triple pattern matches multiple data sources, it must be sent through independent sub-queries to all matching data sources.

The second contribution of DARQ consists in a set of logical (query rewriting) and physical optimizations of the sub-queries produced through the query planning phase. When FILTER expressions are available, the value constraints are reused to replace variables by constants. Moreover, based on the rules proposed in [Pérez et al., 2006, Pérez et al., 2009], the query optimizer merges several basic graph patterns forming the initial SPARQL query. Finally, when possible, value constraints from the FILTER expressions, are incorporated to the sub-queries to avoid transferring useless intermediate results. The physical optimizations consist in finding the query plan that will reduce the size of transferred data, thus achieving faster global querying. The DARQ optimizer is based on a dynamic programming algorithm to find the optimal query plan. The DARQ engine implements both a nested-loop join and a bind join [Haas et al., 1997]. While the nested-loop join naively iterates over bindings provided by the outer relation with bindings provided by the inner relation, the bind join aims at exploiting already known bindings to ultimately replace join variables by their already known values. Finally, sub-queries containing bound join variables are sent multiple times, and allow reducing the size of transferred results.

SPLENDID [Görlitz and Staab, 2011] is a query optimization strategy aiming at transparently federating distributed SPARQL endpoints, by exploiting statistical data describing their content. Statistical data are provided by the Vocabulary of Interlinked Data (VoID) first introduced in [Alexander *et al.*, 2009]. VoID has been designed from both the data producer and the data consumer perspectives, and aims at describing the content of a dataset (a set of RDF triples published through a single provider), its relation to other datasets (interlinking information) and the vocabularies used in the dataset (RDF-S/OWL classes or properties). SPLENDID relies on two main components, namely the index manager and the query optimizer.

The index manager is responsible for associating each VoID description to the corresponding SPARQL endpoint through an index structure. Basically, two indices are provided, the first one describing the cardinality of a given predicate (or RDFS/OWL property) for a given endpoint, the second one describing the cardinality of a given type (or RDFS/OWL class) for a given endpoint.

VoID descriptions are exploited by the query optimizer to reduce the cost of the overall distributed querying. After the query rewriting step, the optimizer performs (i) data source selection and (ii) join re-ordering optimization. The data source selection consists in, for each triple pattern, retrieving a matching data source for the type/predicate indices. If a predicate is unbound in the triple pattern, all data sources are associated since no information is available from the indices. The source selection is refined for triple patterns having bound variables because they are supposed to be located into a single data source. But VoID descriptions do not cover this particular case. To overcome this issue, SPARQL ASK queries are sent to all data sources to determine which one(s) actually host(s) the triple. Indeed, the result of the ASK query (true/false) indicates if a matching triple pattern can be found in the target data source. Except in the case of exclusive groups introduced in [Schwarte *et al.*, 2011] and aiming at grouping triple patterns that are exclusively hosted in a single data source, triple patterns are sent individually to the data sources to not miss any result that would be joined across distributed data sources.

Once the data sources have been accurately selected, the query optimizer performs the join re-ordering, based on a dynamic programming algorithm.

At query runtime, two strategies are used to distribute joins: parallel joins and bind joins. While parallel joins consist in sending requests in parallel to the distributed data sources and finally joining locally the results (a strategy adapted to selective joins, leading to small result sets), bind joins exploit the results obtained from the first join operand into the second operand.

SemWIQ [Langegger *et al*., 2008] is a collaborative knowledge-sharing platform, targeting, in particular, e-Science communities which adopt ontologies to semantically describe scientific data. SemWIQ is based on a mediator-wrapper approach for handling data heterogeneity and provides a unified query interface through an extended SPARQL engine. The general approach consists in (i) analyzing a global SPARQL query designed with respect to a global schema (generally modeled through ontologies), (ii) selecting in a data source registry, relevant data sources based on the concepts referred to in the query, then, (iii) a canonical query plan is optimized through a federation optimizer which groups sub-queries through relevant data sources, and finally (iv) the global query plan is executed over possibly wrapped data sources.

The heterogeneity of data sources is addressed through the DR2-Server, for relational data sources. For non-RDF and non-relational data sources, a local wrapper may allow for remote sub-query plans execution by performing native data access and transformations.

The distributed query processing is achieved as follows. SemWIQ relies on a concept-based data integration approach, thus requiring for all data to be described as instances of ontology classes. Based on the data source registry, source content statistics (RDFStats [Langegger and Wöß, 2009]), and declarative rules (JBoss Rules), simplified Basic Graph Patterns (BGPs) are grouped together, in a first step, to be sent towards appropriate data sources through Service operators. Moreover, Service operators may be combined through (i) Union operators, allowing for querying several target data sources for a same BGP, and (ii) Join operators allowing for combining data resulting from several triple pat- terns sharing the same variables. In a second step, the resulting global query plan is re-ordered trough an iterative dynamic programming algorithm, aiming at finding the cheapest query evaluation plan.

SemWIQ has some limitations with regards to the expressivity of SPARQL. Indeed, (i) only SELECT queries are supported, (ii) subjects of triple patterns must be variables, and (iii) the distribution query processing does not support for unknown data types, i.e. ontology classes of data instances must be known statically (asserted data types), or implicitly deduced through description logic constraints.

FedX [Schwarte et al., 2011] is another SPARQL distributed query processor addressing the querying of federated SPARQL endpoints. The originality of FedX is that no assumptions are made over the content of the distributed data source, and no metadata or statistics are needed to perform static sources selection. Similarly to the approaches presented above, and following the general distributed query processing model proposed in [Kossmann, 2000], FedX performs in a first step data sources selection, followed by a set of static join optimizations (triple pattern grouping and join re-ordering). Final dynamic join optimizations are performed to exploit intermediate results at query runtime.

On-demand data sources selection is realized without any a priori knowledge over the content of the federated data sources. Indeed, each triple pattern forming the basic graph pattern is annotated with a relevant data source through the execution of a SPARQL ASK query which populates a local cache preventing from re-executing the ASK when not necessary.

Once annotated with the relevant data source, triple patterns are grouped together when belonging to exclusive groups. This notion is introduced to characterize triple pat- terns that can be matched into a single data source, and thus do not require for a costly distributed join. This grouping technique can drastically reduce the number of remote invocations and thus the number of transferred results through the network.

Then joins are re-ordered through a rule-based join optimizer. The proposed algorithm is based on the variable counting technique [Stocker *et al*., 2008] to evaluate the cost of the joins. This heuristic consists in estimating the number of free variable considering that subject variables are more selective than object variables, themselves more selective than predicate variables.

Finally, joins are optimized at query runtime by exploiting the intermediate results provided by the previously computed joins. The FedX engine is based on an optimized nested-loop join technique. More precisely, in classical bind join techniques presented above, mappings resulting from the left part of the join are pushed individually to the right part of the join leading to a huge number of remote invocations if the amount of intermediate results (mappings) is high. To avoid this issue, the proposed optimization consists in grouping a set of mappings into a single sub-query through SPARQL UNION constructs, thus avoiding numerous remote invocations. The single sub-query finally need to be post-processed to correctly associate the results of this UNION query to the global result graph. FedX additionally provides a parallel implementation through a pool of worker threads, and a pipelining strategy to exploit intermediate results as soon as they are available.

SPARQL-DQP. Buil-Aranda *et al*. propose in [Buil Aranda and Corcho García, 2010] to federate SPARQL queries over a set of SPARQL endpoints based on relational database distributed query processing (DQP) techniques. SPARQL-DQP relies on the transformation of a subset of SPARQL queries to their equivalent SQL queries. The distributed query processing is then supported by the OGSA-DAI and OGSA-DQP [Lynden et al., 2009] framework.

To address multiple RDF data source querying, the SPARQL 1.0 language has been slightly extended (SPARQL-D) through possibly multiple FROM clauses (included into SELECT queries) allowing determining the source endpoint to which the query must be evaluated. After being parsed, SPARQL-D queries are transformed into a set of SQL queries (based on a relational algebra for SPARQL [Cyganiak, 2005]) which forms a logical query plan (LQP). The query plan is optimized and evaluated through OGSA- DQP. Queries can be evaluated through both pipelined and partitioned parallelism. While pipelined parallelism relies on a multi-threaded implementation of iterators, allowing providing tuples as soon as they become available, for directly being processed by the next operator, the partitioned parallelism relies on several distributed nodes being queried in parallel and providing blocks of result tuples finally being merged by the DQP coordinator.

The RDF data publication is realized through a specific OGSA-DAI data resource, implementing the WS-DAIOnt-RDF(S) specification, and allowing for wrapped RDF data to be accessed through OGSA-DAI/OGSA-DQP.

Finally, a set of OGSA-DAI data-centric workflows are generated from the optimized, partitioned, Logical Query Plan (LQP). These workflows are connected through their inputs and outputs and enacted to return SPARQL results from the multiple virtual RDF data sources.

Optimizations for SPARQL 1.1 Federation. Well-designed graph patterns have been introduced in [Pérez *et al*., 2009] to restrict the usage of SPARQL OPTIONAL clauses thus leading to more effective query evaluations. Based on these results, Buil-Aranda *et al*. propose in [Aranda *et al*., 2011] optimizations based on query

rewriting in the context of distributed SPARQL 1.1 Federation query evaluations. The main idea consists in reordering the initial SPARQL query so that the most selective operators are executed first and OPTIONAL clauses – OPTIONAL being the most expensive SPARQL operator [Pérez et al., 2009] – are executed the latest. These optimizations, and a well-designed graph patterns checker, have been implemented in the SPARQL-DQP framework.

Dynamic source discovery. Approaches presented above are characterized as top-down [Ladwig and Tran, 2010] since they rely on a statically fixed set of distributed data sources, for which it is possible to build a priori knowledge on the content of data sources. To adapt to dynamically evolving data sources, whose availability is unpredictable, Ladwig *et al*. propose a hybrid approach, which consists in dynamically ranking sources at query runtime. The bottom-up query evaluation strategy consists in (i) extracting data sources from the query, (ii) starting to evaluate the query while discovering new data sources from intermediate results at runtime, (iii) evaluating the query against the new data sources, and (iii) terminating the evaluation when all possible data sources have been explored. A hybrid strategy consists in refining, at query evaluation time, an optimized query plan based on partial knowledge on the content of data sources. The proposed implementation rely on non-blocking join operators as introduced in [Hartig et al., 2009] and [Ladwig and Tran, 2011]. This approach is particularly promising since it allows for static optimization, while still considering the dynamicity, and the volatility of data sources distributed over the web.

We described above general-purpose semantic data federation approaches. These approaches are mainly focusing on performance issues. They address the scalability challenge for data integration through performance-oriented distributed query processing techniques, and they aim at lowering the impact on end-users for massive semantic data querying.

The SPARQL-DQP approach, does not address transparent semantic federation, and rely on SPARQL Service clauses. These clauses help in implementing distributed query processing when the content of data sources is well partitioned and known at query design time. However, this assumption does not hold in many real use cases, especially in the context of life-science collaborative platforms. This approach is not suitable in the context of dynamic knowledge base federations in which pre-designed SPARQL 1.1 queries must be adapted to take into account the data source availability.

Addressing transparent semantic federation, DARQ, Splendid, SemWiq, Fedx, and Ladwig *et al*., address performance issues and tend to reduce the amount of data communication between the federation querier and the multiple remote data sources. Precise data source selection becomes thus crucial to prevent from unnecessary communications through optimized distributed joins. Whereas Splendid and DARQ are based on *a priori* knowledge on the content of data sources to perform source selection, this task is dynamically achieved through SPARQL ASK queries in the FedX approach, which provides better flexibility and prevents from maintaining additional data source content descriptions. However Ladwig *et al*. propose an interesting hybrid approach in which source selection can be driven by a priori source content descriptions and a dynamic refinement, thus allowing to adapt to the dynamic nature of data sources federated over the web.

With respect to the knowledge challenge for life-science data federation, all these approaches are based on SPARQL distributed querying, thus allowing, in theory, to exploit ontologies and possible inferences from federated query processing. However all approaches suffer from limitations with respect to the expressiveness of the supported SPARQL features. They all only support SELECT queries and may

impose constraints with regards to the supported basic graph patterns. For instance, SemWiq prevents from federating triple patterns with bound subjects, and all instances must be associated to an ontology class.

Towards a better exploitation of domain ontologies through life-science collaborative platforms, it becomes decisive to find a good balance between expressiveness and performance when federating semantic queries at large scale.