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-14-1-v1

Janvier 2014

Bilan de l'année 2013

O. Corby C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, F. Michel, J. Montagnat

Résumé

Ce document résume le travail réalisé et les résultats du projet CrEDIBLE au cours des deux premières années 2012 et 2013. Il conclue avec les orientations de travail envisagées pour 2014.

Table des matières

1 Contexte et motivations

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 de la nécessité de disposer de référentiels 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, et une nécessité, au vu des quantités de données exploitées. 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 et l'interrogation conjointe d'entrepôts de données distants sont donc des éléments clés du développement du secteur de la recherche médicale.

Le projet CrEDIBLE a pour objectif d'étudier l'intégration des données médicales dans des entrepôts distribués de connaissances. Il recouvre (1) la fusion (virtuelle) d'entrepôts physiquement distribués mais devant apparaître pour leurs exploitants comme une entité unique et cohérente ; (2) l'alignement sémantique de sources de données hétérogènes, qui n'ont souvent pas été conçues pour être exploitées conjointement ; et (3) la description d'ensembles de données distribuées, définis par l'intermédiaire de requêtes qui peuvent s'appliquer sur l'ensemble de la fédération. Nous adoptons une approche fondée sur la représentation sémantique des données, afin de faciliter l'alignement de données hétérogènes (médiation) et l'interrogation conjointe de plusieurs entrepôts de données (fédération).

Les principaux verrous scientifiques abordés sont :

- La représentation sémantique des données d'imagerie médicale (recherche clinique, imagerie, traitement des images, marqueurs quantitatifs associés à des structures anatomiques ou processus physiologiques ou physiopathologiques, scores cliniques…) ;
- La gestion de sources de données hétérogènes par des mécanismes de médiation dynamiques fondés sur la sémantique des données ;
- La fédération d'entrepôts distribués par le biais de mécanismes de requêtes applicables à l'ensemble de la fédération (distribution, réécriture) ; et
- La performance des requêtes distribuées.

2 Travail réalisé

Le travail réalisé au cours des deux premières années du projet a consisté en :

- **Réseau scientifique**. Une étude élargie du domaine abordé, en particulier des ressources ontologiques [1], des moteurs d'interrogation de bases de données distribuées [5, R2] et des techniques de médiation [6, R4]. Pour assurer une bonne prise en compte des travaux existants dans ce domaine, construire un réseau scientifique de haut niveau et permettre une bonne diffusion du projet

CrEDIBLE, deux ateliers multidisciplinaires ont été organisés à Sophia Antipolis en 2012 puis en 2013. Ils ont fait 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.

- **Prototype**. KGRAM-DQP (Knowledge Graph Abstract Machine – Distributed Query Processing) [2, 3, 4]: un moteur de recherche dans des entrepôts de données sémantiques distribuées, fondé sur les standards du Web Sémantique (RDF, OWL, SPARQL) et basé sur le moteur KGRAM développé à INRIA[1].
- **Modélisation des données**. Une ontologie générique (indépendante de tout domaine scientifique) – DataTop – fournissant une référence sémantique pour rendre compte du contenu et du contexte de production des données. Cette année, les travaux de modélisation ont essentiellement porté sur une définition des résultats de mesures issues d'observations scientifiques [R5].

2.1 Ateliers multidisciplinaires.

Les ateliers de travail multidisciplinaires ses sont déroulés du 15 au 17 octobre 2012[2] et du 2 au 4 octobre 2013[3] respectivement. Ils avaient 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. Chaque atelier a fait intervenir une vingtaine d'orateurs et a réuni environ 35 participants. Le résumé et les conclusions de chaque atelier ont été restitués dans des rapports du projet [R1, R6].

Le premier atelier était constitué de 4 sessions portant sur les techniques d'intégration de données et les outils existants ; les ontologies et la modélisation sémantique ; les modèles de représentation de données ; et le raisonnement et les workflows sémantiques.

Tenant compte de l'expérience acquise en 2012, l'atelier de 2013 a laissé plus de place aux interactions entre les participants par l'intermédiaire de panels de discussion associés à chaque session. Il a été structuré en 5 sessions portant sur les entrepôts de données et leur réutilisation en recherche clinique ; les ontologies biomédicales ; la médiation de données ; la fédération de données ; et les graphes de données et le raisonnement.

Ces ateliers ont reçu un accueil très favorable de la communauté internationale, avec beaucoup d'enthousiasme de la part des contributeurs (des orateurs sont venus de toute l'Europe et même d'Amérique du Nord et du Sud) et des retours très positifs. Le mode d'organisation, avec invitation des orateurs, a permis de concevoir des sessions mieux ciblées et plus homogènes que dans la plupart des ateliers traditionnels. En outre, le spectre de discipline couvert a souvent été jugé très favorablement, un tel éventail d'expertise étant particulièrement difficile à réunir traditionnellement.

[1] Code open-source KGRAM disponible depuis http://www-sop.inria.fr/edelweiss/software/corese/kgram

[2] Atelier 2012 et programme de travail : https://credible.i3s.unice.fr/doku.php?id=atelier_15-17_octobre

[3] Atelier 2013 et programme de travail : https://credible.i3s.unice.fr/doku.php?id=2013_workshop

2.2 Etudes bibliographiques

Le projet CrEDIBLE s'est intéressé en particulier à l'étude des ontologies utilisées pour la représentation des connaissances en imagerie biomédicale, aux techniques d'interrogation de bases de connaissances distribuées, et aux techniques de médiation pour l'alignement de données hétérogènes.

Ontologies médicales. En s'appuyant à la fois sur l'expérience du projet ANR NeuroLOG concernant la conceptualisation d'instruments de mesures utilisés en neurosciences[4] [1] et les présentations de travaux récents faites lors des ateliers multidisciplinaires CrEDIBLE [R2, R4], un cadre ontologique a été défini pour rendre compte de la sémantique des données d'observation.

Interrogation de bases de connaissances distribuées. L'interrogation d'entrepôts sémantiques distribués est une problématique qui suscite beaucoup d'intérêt. On distingue la répartition "verticale" des données, lorsque des entrepôts différents contiennent des entités différentes (mais qui peuvent néanmoins être reliées lorsque leur sémantique est connue) et la répartition "horizontale", lorsque des entités de même nature sont réparties dans plusieurs bases. Les entrepôts de données biomédicales sont concernés par les deux types de partitionnement 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 une 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 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 (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, souvent au détriment de la richesse du langage de requête ou en ignorant la possibilité de répartition horizontale des données. Les technologies du Web sémantique apportent cependant une grande flexibilité par la richesse du langage d'interrogation SPARQL et la possibilité d'inférer des connaissances à travers l'application de règles déductives qui sont déterminantes dans les applications médicales. Une étude bibliographique détaillée des principales approches documentées dans la littérature a été réalisée dans le cadre du projet CrEDIBLE [5, R2] et ses conclusions nous amènent à considérer la conformité avec le standard SPARQL dans un cadre distribué.

Médiation de schémas de données et d'interface d'interrogation hétérogènes. Les entrepôts de données biomédicales existant font très majoritairement appel à des schémas de données relationnels. L'intégration de telles bases dans un système de fédération de données basé sur les technologies du web sémantique et un modèle ontologique pivot nécessite l'explicitation de la sémantique des données relationnelles et leur alignement sur le modèle de référence. De nombreuses techniques de transformation des données relationnelles en données RDF ont vu le

[4] Le projet ANR TechLog NeuroLOG (2007-2010), précurseur du projet CrEDIBLE, a abouti à la définitions de trois modules ontologiques pour la modélisation des ensembles de données IRM (http://bioportal.bioontology.org/ontologies/ONL-MR-DA), des instruments de mesure des désordres mentaux ou tests cliniques (http://bioportal.bioontology.org/ontologies/ONL-MSA) et des traitements d'images (http://bioportal.bioontology.org/ontologies/ONL-DP).

jour ces dernières années, en particulier le langage standard R2RML du W3C. Il ne s'agit cependant que d'un langage, qui ne décrit pas le processus de transformation et laisse beaucoup de questions ouvertes quant à son opérationnalisation (par exemple en utilisant des techniques de réécriture à la volée ou au contraire en procédant à des transformations globales des entrepôts sources). Une étude détaillée du standard R2RML et des outils l'implantant entièrement ou en partie a été réalisée [6, R4]. Un travail d'opérationnalisation s'appuyant sur les conclusions de cette étude est en cours et un élargissement à d'autres types d'entrepôt (XML ou NoSQL) débute avec un stage de master.

2.3 Prototype

Un prototype de moteur de recherche dans des entrepôts sémantiques distribués a été implanté. Il est basé sur le logiciel KGRAM (Knowledge Graph Abstract Machine) développé à 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 [2, 3, 4]. 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].

2.4 Ontologie dédiée aux données : DataTop

Un cadre ontologique organisé autour de DOLCE-CORE, une nouvelle version de l'ontologie fondatrice DOLCE développée au LOA (ISTC-CNR, Trento, Italie), a été défini [R5]. 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 aux 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.

3 Autres Résultats

Collaborations. La visibilité du projet CrEDIBLE a conduit à plusieurs invitations des partenaires dans différents colloques scientifiques (Réseau "Intégration de sources/masses de données hétérogènes" de l'INRA[5], ateliers des projets ANR AnaEE[6] et BIOMIST). Une nouvelle collaboration avec l'Université de Laval au Québec sur la modélisation ontologique de la maladie d'Alzheimer est également en train de débuter.

Soumissions de projets. Le projet CrEDIBLE est à l'origine d'une proposition de projet ANR sur sa thématique, soumise à l'appel CONTINT 2013 (projet non retenu malgré une évaluation positive et en cours de resoumission sur l'appel générique 2014 dans les défis « société de l'information et de la communication » et « santé et bien-être »). Par ailleurs, deux partenaires de CrEDIBLE ont répondu ensemble à un appel à projets européen dans le cadre du projet Flagship 'Human Brain Project'

[5] https://www6.jouy.inra.fr/metarisk/Research-Unit/Knowledge-Engineering/Reseau-MIA/Reunion-du-28-11-2013

[6] http://www.anaee-s.fr/spip.php?article49

(intitulé « Rich and Simple Data Sharing »). Enfin, un second projet européen de type COST (intitulé 'European Brain Imaging Network for Psychosis') est en cours de soumission.

Thèse. Alban Gaignard, encadré par Johan Montagnat, a soutenu sa thèse [4] portant sur le volet requêtes distribuées de CrEDIBLE en mars 2013. Un stage de M2 sur la médiation de bases de données relationnelles ou NoSQL est également en cours. Enfin, la thèse de Nadia Cerezo [7], encadrée par Johan Montagnat et soutenue en décembre 2013, a exploité l'ontologie développée dans le cadre de CrEDIBLE pour la modélisation de chaines de traitement d'images médicale.

4 Objectifs 2014

Les objectifs de l'année 2014 concernent l'optimisation du moteur KGRAM-DQP de requêtes distribuées, l'intégration de techniques de médiation pour des sources hétérogènes et l'enrichissement de l'ontologie DataTop.

Optimisation du moteur de requêtes KGRAM-DQP. Le prototype implanté et l'étude des différents outils existants pour la réalisation de requêtes sémantiques sur des entrepôts distribués ont permis d'identifier précisément les défauts des stratégies d'optimisation existantes, notamment lorsqu'il s'agit de prendre en compte simultanément la répartition verticale et horizontale des données (les heuristiques d'optimisation pour traiter ces deux cas différant et étant parfois contradictoires). Une restructuration du code de KGRAM est en cours pour assurer un fonctionnement complètement asynchrone du moteur de requêtes et permettre la mise en place de stratégies d'optimisation alternatives en fonction du contenu des sources de données.

Médiation de sources hétérogènes. Une implantation du standard R2RML est en cours pour permettre l'intégration de sources relationnelles dans la fédération de données. En outre, les autres modèles de base de données non relationnels sont en cours d'étude pour envisager un élargissement à des sources différentes, notamment XML et NoSQL, à travers un langage aussi proche que possible de R2RML.

Enrichissement de l'ontologie DataTop. Un travail concernant l'ontologie des valeurs (quantitatives et qualitatives), des échelles et des unités de mesure, est en cours en collaboration avec le LOA (Trento, Italie). La distribution (physique et temporelle) des données ainsi que la structuration des données (composition, collections) seront étudiées pour compléter DataTop.

5 Publications

[1] 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), septembre 2012. (orateur invité).

[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, octobre 2012.

[3] 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, décembre 2012.

[4] A. Gaignard. "Distributed knowledge sharing and production through collaborative e-Science platforms", *Thèse de l'Université de Nice Sophia Antipolis*, 254 pages, France, mars 2013.

[5] A. Gaignard, O. Corby, C. Faron Zucker et J. Montagnat. "Provenance-based summarisation of Life-Science e-experiments" soumis à *Journal of Web Semantics*. juillet 2013.

[6] F. Michel, J. Montagnat et C. Faron-Zucker. "A survey of RDB to RDF translation approaches and tools", soumis à *Journal of Web Semantics*, novembre 2013.

[7] N. Cerezo, "Conceptual Workflows", *Thèse de l'Université de Nice Sophia Antipolis*, 213 pages, France, décembre 2013.

6 Rapports de recherche

[R1] O. Corby, C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, J. Montagnat. "CrEDIBLE multi-disciplinary workshop (2012)". Rapport CrEDIBLE-12-1-v1. novembre 2012.

[R2] C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, J. Montagnat. "Bilan de l'année 2012". Rapport CrEDIBLE-12-2-v1. décembre 2012.

[R3] C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, J. Montagnat. "Résumé d'activité 2012". Rapport CrEDIBLE-12-3-v1. décembre 2012.

[R4] F. Michel, J. Montagnat et C. Faron-Zucker. "A survey of RDB to RDF translation approaches and tools", *rapport I3S 2013-04-FR*, 25 pages, laboratoire I3S, Sophia Antipolis, novembre 2013.

[R5] B. Gibaud et G. Kassel. "Sémantique des données de l'observation : une approche ontologique". *Rapport numéro CrEDIBLE-13-1-v1*. décembre 2013.

[R6] O. Corby C. Faron Zucker, A. Gaignard, B. Gibaud, G. Kassel, F. Michel et J. Montagnat. "CrEDIBLE multi-disciplinary workshop", *Rapport numéro CrEDIBLE-14-1-v1*. janvier 2014.