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-13-3-v1
Décembre 2013

CrEDIBLE Multi-disciplinary workshop

Sophia Antipolis, 2-4 October 2013

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

Summary

This document summarises the output of the CrEDIBLE multi-disciplinary workshop organized in Sophia Antipolis in October 2-4, 2013. This second issue of the workshop aimed at gathering scientists from all disciplines involved in the set up of distributed and heterogeneous medical image data sharing systems, to provide an overview of this broad and complex area, to assess the state-of-the-art methods and technologies addressing it, and to discuss the open scientific questions raised.

Table of content

1 Workshop objectives

The CrEDIBLE project organized 3 multi-disciplinary working days in October 2-4, 2013 in Sophia Antipolis (France) where experts were invited to discuss their approaches for biomedical data management. The aim was to gather scientists from all disciplines involved in the set up of distributed and heterogeneous medical image data sharing systems, to provide an overview of this broad and complex area, to assess the state-of-the-art methods and technologies addressing it, and to discuss the open scientific questions raised.

1.1 Summary

This multi-disciplinary workshop featured 5 thematic sessions over the 3 workshop days. Each session included 30 minutes long oral talks from invited speakers and a panel discussion with all session speakers where the audience was invited to interact and discuss the session topic, challenges and perspectives. Two moderators from the CrEDIBLE consortium led each panel discussion. A list of challenging questions to be addressed during panel discussions were prepared and distributed to the speakers ahead of the meeting.

The five themes addressed were clinical data repositories, biomedical ontologies, data mediation, data federation and graphs and reasoning:

- **Data repositories for secondary use of clinical and research data.** Reporting on concrete experiences in developing systems that gather or index data to be shared and reused in biomedical research projects, with particular focus on user requirements, current technology limitations and future expectations.
- **Biomedical ontologies.** Discussing ontologies for modeling scientific observations and measurements data (designed to facilitate the sharing and reuse of scientific data).
- **Data mediation.** Data mediation is needed to federate heterogeneous data stores and perform semantic alignment of different data models onto a reference model. In the biomedical domain, both different kind of data (e.g. biological samples, medical images, clinical records…) and different data models for the same king of data are considered.
- **Data federation.** Biomedical data stores may be partitioned vertically (different stores containing different, complementary kind of information) and/or horizontally (different stores containing similar data entities). The simultaneous federation of horizontally and vertically partitioned data stores is particularly challenging and it has an impact on query optimization strategies and achievable performance.
- **Graphs and reasoning.** Semantic Web technologies are widely adopted to represent, align and query heterogeneous data stores. Knowledge graphs can also be used to infer new information from the base of known facts. Data distribution and the scale of the data federation often challenge this reasoning capability though.

1.2 Programme

The detailed programme of each session and the list of questions addressed during the panel discussions is given below. Presentation slides are available from the workshop web page.

Session 1, Data repositories for secondary use of clinical and research data

- **M. Cuggia** (U. Rennes 1, France): Secondary use of Clinical Data for Medical Research

- **M. Martone** (UCSD, USA): Experience with the development and operation of the Neuroscience Information Framework (NIF) portal
- **S. Villata** (INRIA, FR): Applying open data provenance and licensing to biomedical data
- **K. Belhajjame** (U. Manchester, UK): Research Objects: Preserving Scientific Workflows and Their Provenance
- **C. Marion** (Kitware): Visualization and analysis of medical data through the Internet
- **Panel discussion:**
    - Data indexing: How to meet the expectations of researchers in terms of precision of the vocabulary?
    - Data federation: What level of data federation is required? What are the data sources to federate? What are the data models in use?
    - Data provenance: Are detailed models of provenance or information summary more relevant for data reuse?

Session 2: Biomedical ontologies

- **B. Gibaud** (INSERM, France) and Gilles Kassel (U. Picardie, France): Observation data semantics: an ontological approach
- **C. Masolo** (LOA, ISTC-CNR, Trento, Italy): Quality-spaces: problematic aspects
- **W. Kuhn** (U. Münster): Ontology of observations in space and time
- **M. Martone** (UCSD, USA): Experience of indexing brain research related measurements with NIFSTD
- **Panel discussion:**
    - How to model related entities (observed entity, measured quality, measurement results, units of measurements) and relationships?
    - Relation and compatibility with foundational ontologies such as DOLCE, DOLCE-CORE and BFO? With existing ontologies of qualities?
    - How to model complex observation data such as images?
    - How to model time varying phenomena?

Session 3: Data mediation

- **N. Lopes** (National University of Ireland Galway), XML data mediation using XSPARQL
- **J. Euzenat** (INRIA, France), Data mediation in SPARQL from alignments
- **M. Vincent** (Logilab, France), BRAINOMICS: A management system for exploring and merging heterogeneous brain mapping data based on CubicWeb
- **Panel discussion:**
    - Taxonomies or ontologies can be used as reference model. Are they most appropriate reference model?
    - Is SPARQL the most appropriate query language to access data in heterogeneous databases?
    - How to mediate various data sources? Statically (ETL), transforming data sources periodically? Or dynamically, on the fly?
    - How to ensure access control in an heterogeneous deployment?

Session 4: Data federation

- **A. Schwarte** (fluid Operations AG, Germany), FedX: A framework for efficiently evaluating SPARQL queries in a federated environment
- **M.-E. Vidal** (U. Simón Bolí-var, Caracas, Venezuela), On the Efficiency and Effectiveness of Federated Semantic Data Management - ANAPSID An Adaptive Approach

- **P. Molli** (U. Nantes, France), SemLav: Local-As-View mediation for SPARQL Queries
- **F. Paulus** (SemSoft, France), Data federation tools at SemSoft
- **Panel discussion:**
    - Is SPARQL the most appropriate language for distributed querying? What is the trade-off between expressiveness and performance?
    - What is the performance impact of distribution? Gain of parallel execution of queries vs network overhead, especially when deploying over a WAN?
    - How scalable are the different methods proposed? To what scale have they been tested?
    - What is the impact of low reliability? Can queries be partially processed in case of communication failures with some data stores? Can end-users be notified on the kind of potentially missing information?

Session 5: Graphs and reasoning

- **J. Urbani** (Vrije Universiteit, Amsterdam), Forward versus Backward: Two approaches for web-scale reasoning
- **M.-L. Mugnier** (LIRMM, Montpellier, France), Ontology-based Query Answering with Existential Rules
- **O. Curé** (Université Marne La Vallée, France), RDF triple stores and indexation
- **Panel discussion:**
    - How to process large RDF graphs? (Storage in databases, scalability of graph processing algorithms, graphs indexing.)
    - How can semantics described in ontologies be used to interpret RDF data? Trade-off to be found between the amount of data to process and the reasoning capabilities of the system?
    - Other scalability opportunities when addressing data querying: top-k query answer algorithms, probabilistic algorithms?

The slides associated to all talks are available online from the CrEDIBLE workshop website. The output of each session is summarized below.

2 First session: Data repositories for secondary use of clinical and research data

The first session "Data repositories for secondary use of clinical and research data" had two major objectives:

- To describe significant experiences of deployment of such data repositories, and draw lessons about their level of maturity in terms of how they meet user communities' expectations regarding, e.g., their function, performance, provisions concerning access control etc.
- To share concrete situations that may be referred to in the following of the workshop, when discussing, e.g., the capabilities of new technology or new approaches for querying, and mediating data at a large scale.

The first presentation entitled "**Secondary use of clinical data for medical research**" was given by Marc Cuggia, Professor of Medical Informatics at the University of Rennes (France). He first highlighted the key role of informatics in translational medicine, to connect basic science, clinical research and public health research into a synergistic whole. He introduced a basic distinction between health information systems dedicated to care delivery, that are primarily patient-centred and research information systems oriented to secondary use (for research or evaluation of practice), that are organized to analyze groups of patients or populations.

He introduced ASTEC, a system aiming at facilitating the recruitment of patients for inclusion into clinical trials. He then presented ROOGLE, a system aiming at indexing large corpora of clinical data for secondary use, and offering querying capabilities with both structured data search and free text search. He then discussed some of the issues raised by pooling data from multiple institutions. He illustrated these issues with achievements from two European projects. The first, EHR4CR (FP7/IMI), focuses on recruitment for clinical trials, and the second, DEBUGIT, is exploring how clinical data repositories might be used to assist the assessment of antibiotics resistance evolution in multiple European healthcare institutions.
As a conclusion, Marc mentioned interesting perspectives in associating NLP methods and ontology-based indexing.

The second presentation was given by Maryann Martone, Professor at University of California San Diego (USA) and PI of the Neuroscience Information Framework (NIF) project. Her talk focused on the setting up of the **NIF portal**, a platform for indexing a broad spectrum of resources such as web sites, databases, scientific literature, etc.
As an introduction, Maryann insisted very much on the basic principles that guided NIF along the 5-year deployment, especially the need to cope with the current state of these resources (rather than trying to change it) and the constraint to cover a very broad field and to populate the system rapidly.
She then briefly introduced the NIF Standard ontology (NIFSTD), a modular application ontology covering multiple structural scales and based on existing ontological resources. She explained the complementarity between NIFSTD and NeuroLex, a semantic wiki (searched by Google) facilitating the direct contribution from neuroscientists (hardly feasible for devolping or maintaining NIFSTD). Maryann introduced also the NIF data federation, supported by DISCO, a software resource developed at Yale University and facilitating the rapid integration of resources.
The following of the talk focused on NIF users (primarily neuroscientists), main uses and software resources for providing resource growth statistics and ensuring that resources are still "alive". As a conclusion, Maryann stressed how challenging it was to keep resources up to date.

The third presentation "**Applying open data provenance and licensing to biomedical data**" was given by Serena Villata, from INRIA WIMMICS at Sophia Antipolis (France). She first introduced the DATALIFT French project, which aimed at "accelerating the lifting from raw data to linked public data".
In a first part her presentation focused on SHI3LD, a resource for managing access policies to data for which there exist restrictions. SHI3LD applies to data represented using semantic Web standards, organized in arbitrarily complex graphs and can be plugged to any RDF store offering a SPARQL 1.1 endpoint. The SHI3LD vocabulary allows expressing access conditions as well as characterizing the context of the required access (device, user, environment).
The second part dealt with licenses in a web of data, and how licenses could be selected, assessed (for determining compatibility) and composed, and she mentioned some of the related issues and challenges.

The forth presentation "**Research objects: preserving scientific workflows and their provenance**" was given by Khalid Belhajjame". Khalid has worked several years in the University of Manchester (UK) and is now Professor at University Paris Dauphine (France).
Khalid first stressed the importance of reproducibility in science, and introduced 'Research Objects' as a key ingredient of reproducibility. Among the items that can be aggregated in Research Objects, emphasis was put on data processing workflows, other artefacts involved in the computation processes, the data being processed or produced, their detailed provenance, as well as the computing resources involved.

In a second part, he stressed the need for summarizing workflows at a more abstract level (than, e.g., the required level for managing programs' invocation or provenance) and he described various strategies to do it, by elimination or by collapse.

The fifth and last presentation "**Visualization and analysis of medical data through the internet**" was given by Charles Marion, from KITWARE in Lyon (France). His presentation focused on MIDAS, an open source software environment to share medical imaging data across the Web. Emphasis was put on metadata and search capabilities. Regarding metadata, Kitware adopted a generic key-value approach and developed modules to automatically extract metadata from DICOM images. Both SQL queries and free-text search based on APACHE Solr (Lucene) are available for data querying. The use of MIDAS was illustrated with iDASH, a project for sharing medical imaging datasets for research and 3D Slicer DataStore, a resource for sharing data produced/used by the 3D Slicer software package.

The **panel discussion** addressed difficulties to adopt a single vocabulary, and respective advantages and drawbacks of a top-down (prescriptive approach) versus bottom up (vocabularies actually in use). Another difficulty is to find the appropriate balance in terms of complexity, to deliver sufficient precision in denoting shared entities, while staying general enough to be understood consistently across broad and potentially heterogeneous user communities. Whereas it is clear that one should start (as NIF did) with a pragmatic non-intrusive bottom-up approach, it seems necessary to progressively associate data providers in the alignment to a common model (to be sure that data semantics are not misunderstood). Discussions also dealt with interesting opportunities for actual sharing of imaging datasets, which is not currently achieved in NIF, e.g. in the context of the Human Connectome Project (HCP) for comparing the performance of various methods of analysis of MR diffusion data. This raises the issue of data licensing since access to HCP data (as well as other similar repositories) is currently restricted.
The discussion also mentioned the progress of open data in science, but additional incentives would be needed to convince researchers to publish their data more systematically (such as data citation).

3 Second session: Biomedical ontologies

The second session aimed at discussing ontological models for modeling observations and measurements, which obviously play a key role in most scientific activities.

The first presentation "**Observation data semantics: an ontological approach**" was given by Bernard Gibaud, Inserm researcher in LTSI in Rennes (France) and Gilles Kassel, Professor in University of Picardie (France).
The presentation aimed at presenting an ontological framework called 'DataTop', currently under development for representing observations, and inspired by the international vocabulary of metrology (BIMP Joint Committee for Guides in Metrology 2012). This framework aims at covering a wide range of observations, quantitative or qualitative, atomic or composite (such as images), resulting from direct observations or derived from observation data. DataTop is built on top of DOLCE-CORE, a new release of the DOLCE foundational ontology, extended with a set of generic ontology modules providing entities such as Actions, Artifacts, Inscriptions, Expressions and Conceptualizations.
The presenters recalled how the qualities of any entities were modeled in DOLCE-CORE and introduced a few exemplary use cases from the neuroimaging domain. They then focused on the introduction of 'Values' ('quantitative value' or 'qualitative value') and decribed how they are mapped to DOLCE-CORE's Qualities and Regions (using rValue and qValue properties).

The second presentation "**Quality-spaces: problematic aspects**" was given by Claudio Masolo from the LOA in Trento (Italy). The presentation was organized in two main parts, focusing on qualities in DOLCE and DOLCE-CORE, on the one hand, and measurement, on the other hand.

Claudio first recalled how quality spaces were represented in DOLCE. He mentioned the constraint of a bijective relationship between quality kinds and quality spaces. He mentioned that DOLCE-CORE introduces flexibility in the modeling, with qualities being located in possibly several quality spaces.

He then focused on the question whether 'being two meters' is a property? He distinguished different options. In the first, this is considered an abstraction of the different qualities (width, length, etc.). In the second 'being two meters' is understood as 'being two meters long', which means that qualities such as width, height, etc, have a quality ('length') which value is 'being two meters long'. In the third option, the meter can be used to measure height, width, etc, based on the measurement procedure.

In the second part of his presentation Claudio analyzed the three key relationships involved in measurement, namely evaluation along a quality type, classification along this quality type and symbolisation. He related the 'Quality' / 'Quality along Quality type'/ 'Symbol' triplet to the semiotic triangle Referent / Concept /Symbol.

The third presentation "**Ontology of observations in space and time**" was given by Werner Kuhn from The University of Münster (Germany).

In his introduction Werner underlined that existing specifications were very much focused on device technology and syntax, rather than semantics of observations (e.g. what can be observed? how do observations relate to reality?). He considered that this is really an issue when one wishes to share interpretations of observations, or aggregate multiple observations gathered from multiple observers or observation devices.

He then recalled how observations were structured in DOLCE and he presented salient aspects of FOOM, a 'Functional Ontology of Observation and Measurement' based on DOLCE and written in Haskell, used as an algebraic specification language. He noted that neither the choice of qualities, nor their bearers, nor their link to stimuli or assigned values are mind-independent, but rather rely on human constructions. He also noted that entities such as stimuli and qualia ("that are the glue of the ontology") can be abstracted away in final representation. The latter is expressed by an agent who symbolizes the result of the observation.

Werner mentioned briefly how images could be managed, either as an observation of a (single) field, or as an aggregate of observations. He closed his talk with a number of open aspects related to observation, e.g. granularity, accuracy, trust and perception action cycles.

The fourth presentation "**Experience of indexing brain research related measurements with NIFSTD**" was given by Maryann Martone, Professor at University of California San Diego (USA) and PI of the Neuroscience Information Framework (NIF).

Maryann's presentation was composed of two main parts. In a first part, she recalled the principles underlying NIFSTD, the standard ontology supporting the Neuroscience Information Framework (NIF): modular structure, components built according to OBO best practices, representation in OWL 2, alignment to the Basic Formal Ontology (BFO), single coverage of a sub-domain, use of additional cross-module relationships, complementarities with the NeuroLex wiki.

The second part of her talk was focused on representing measurements. She explained that data retrieved from different sources could actually be represented in various and inconsistent ways. In this context, NIF "translates" common concepts, based on the knowledge embedded in the ontology. However, only a small percentage of NIF queries contain references to quantities, yet (around 1%).

Nevertheless there is a clearly identified need to translate quantitative representations into qualitative ones, and to infer equivalence between, e.g., phenotype statements. Therefore, it is planned to include in NIF new ontology modules for better representing measurements, e.g. organism stage, scores resulting from neuropsychological and neurological tests and assessments.

Discussions came back on the difficulties to adopt a single vocabulary, and on the practical added value of foundational ontologies and upper ontologies such as ontologies of observations. It was agreed that the latter helps clarifying the overall conceptualization, but should not necessarily be part of any implementation. This should depend of actual needs in terms of reasoning in each application context.

4 Third session: Data mediation

Mediation refers to the ability to overcome the mismatch between knowledge formalized in heterogeneous data sources, which were often designed independently from each other. Federation of different data sources is hampered by mismatches in representation languages, terminologies used, discrepancies in the way things are modelled, varying scopes and points of views. The session on Data Mediation addressed top-down and bottom-up techniques to reconcile resources expressed through different data models.

Three talks were given during this session. The first two presented existing software, namely XSPARQL and Bainomics, while the third talk addressed more general considerations about ontology alignments.

XSPARQL. Nuno Lopes presented XSPARQL[1], a query language combining XQuery (a language designed to query XML data) with SPARQL and SQL for transformations between RDF, relational and XML data, in any direction. XSPARQL merges SPARQL and SQL components into XQuery FLWOR expressions (For-Let-Where-Order-Return). It is typically designed to address issues such as extracting RDF data out of existing (X)HTML Web pages or relational databases, allowing an RDF-based client software to communicate with XML-based Web services, or enriching an RDF graph with deduction rules described as RDF-to-RDF mapping. Any number of any type of data source (RDF, XML, relational) can be queried and mediated simultaneously to produce an RDF mash-up.

Being based on XQuery, XSPARQL leverages the advantages of its expressiveness such as the scripting of commands and arbitrary nesting of expressions. An appropriate XSPARQL document is able to interpret a W3C R2RML mapping document. In turn, XSPARQL materializes the RDF data by querying the relational database. Annotated RDF can be used in XSPARQL to provide fine-grained Access Control over RDF data (at the level of each triple). Several use cases were shortly presented: Inparanoid, Cloudsapces & Sindice, Logainm.

Brainomics. Vincent Michel, from Logilab[2] company, presented Brainomics[3], an open-source solution for the management of heterogeneous neuroimaging and genomic data. Brainomics adopts a data-warehouse approach, in which all data (neuro-images, genome, result of behavioural tests) from multiple studies are imported into a common repository thus aligning data on a common data model.

Brainomics is based on Cubicweb[4] (also developed by Logilab), an open source framework to help developing semantic Web-enabled data management applications. A data model is defined through a Cubicweb schema: an entity-

[1] http://xsparql.deri.org/

[2] http://www.logilab.fr/

[3] http://www.brainomics.net/demo

[4] http://www.cubicweb.org/

relationship model, that allows for the definition of constraints and embedded fine-grained security rules. Additional business logic can be described along with views used to output data in any format (HTML, RDF, XML, JSON, binary...). Any information is stored in a relational database. The choice of not using a triple store comes from the need for performance and stability and historical reasons (only few triple stores were available at the time Cubicweb was developed).

The expressiveness of the data model language is very similar to that of RDF, and allows linking with existing taxonomies or ontologies. Cubicweb comes with existing data models reflecting common ontologies. Those can be used as is or extended and customized as needed. Data can be imported from/exported to RDF through the description of mappings. Two data input methods are provided: either as bulk loading (similarly to the data warehouse approach), or through the periodic import of data typically in the form of RSS feeds or streamed RDF data.

Data is queried using RQL, the Relational Query Language. RQL is similar to SPARQL 1.0 but is designed to be significantly user-friendlier. Data types can be inferred using subsumption relations of the data model. The rationale behind RQL is to provide end users with the ability to explore the model and express complex queries, rather than offering them easy-to-use but limited query forms. RQL also supports the querying of federated databases using a clause similar to the SPARQL 1.1 SERVICE clause.

Data mediation in SPARQL from alignments. The talk from Jérôme Euzenat described how ontology alignment techniques can be used to reconcile different data models and perform data mediation. Ontology reconciliation requires finding the correspondences between entities (e.g., classes, properties) occurring in different ontologies. Alignments are the declarative expression of a set of such correspondences. The ontology reconciliation process takes place in three steps: (i) an ontology matcher determines an alignment; (ii) the alignment is used to generate a mediator that can take several forms: data transformation engine (translator), query rewriter (mediator), generator of links between entities, ontology merger; (iii) the processor is then applied to the data. W3C R2RML mappings comply with this description: an R2RML document describes the alignments between a relational model and an ontology, from this alignment we can generate two types of processors: either a translation engine that translates relational data into an equivalent RDF graph, or a mediator that rewrites SPARQL queries into SQL queries and converts SQL results into SPARQL results.

Alignments are described as schema-level correspondences (equivalence of classes or properties) possibly involving the expression of complex constraints. It is also often necessary to describe instance-level alignments to describe links between individuals (equality of individuals, similarity of property values, creation of links between individuals using object properties), or data conversion (for example due to varying units used in measures).

SPARQL may be used to perform alignments either at schema-level or instance-level. Nevertheless, expressing some complex constraints such as similarity measures (e.g. approximate equality of strings) will require more expressive languages, such as Silk[5]. The Alignment API[6] is a framework to build, manage and share ontology alignments, described using the EODAL[7] language (Expressive and Declarative Ontology Alignment Language).

Synthesis. A large variety of solutions can apply to mediate and query heterogeneous data sources. They roughly fall into two families: the transformation of

[5] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/

[6] http://alignapi.gforge.inria.fr/

[7] http://alignapi.gforge.inria.fr/edoal.html

data from a native model into a pivot format (like RDF for XSPARQL, a proprietary entity-relationship model for CubicWeb), and ontology alignments methods. Both approaches may be used complementarily. The use of RDF as a representation format for data mediation, as a pivot format, marks the consensus among the approaches presented. Its flexibility is the result of several key elements:

- RDF is self-describing and schema-less, but at the same time it allows for vocabulary reuse.
- Its graph-based model is flexible enough to represent any kind of data. Consequently, combining different data sets by merging RDF is easy.
- The sharing of RDF is very much facilitated by the underlying Web technologies (HTTP, URIs), and the subsequent Linked Data paradigm.

The use of data transformation vs. ontology alignment methods was discussed during the panel discussion of this session. In particular, with regards to relational databases, lots of different combinations may be considered:

- Relational data can be translated into an equivalent RDF graph (also called the materialisation approach), or it can be queried through a mediator that rewrites SPARQL queries into SQL queries and converts SQL results into SPARQL results. The second option will be better fitted in context where dynamicity and data freshness are an issue.
- The RDF data generated from the relational data can result from a simple Direct Mapping or a complex set of alignments using a mapping description language such as R2RML.
- The RDF data generated from the relational data can be further lifted using ontology alignments techniques such as those described by Jérôme. Additional instance-level alignments can be used to describe links between individuals.
- Provided those different ways to mediate heterogeneous data sources, we can think of different scenarios of federated querying by combining the alignment techniques (topic of the next session):
  - The RDF materialisation of several data sources may be merged into a single graph.
  - A query federator may simultaneously query materialized RDF data and native data using a mediator that rewrites SPARQL queries into the native query language.
  - Ontology alignments can help aligning the RDF representation of different data sources, either using the data materialisation or query rewriting approach.
  - Additional instance-level alignments can be used to describe links between individuals.

5 Fourth session: Data federation

This session addressed data federation issues commonly faced when setting-up distributed and heterogeneous medical data sharing systems. The first two presentations focused on distributed query processing and introduced the FedX and ANAPSID federation engines. The last two presentations focused on mediating heterogeneous data sources through two *Local-As-View* approaches, namely SemLAV and Agreggo.

FedX. Andreas Schwarte, from fluid Operations AG (Germany), presented "*FedX, A framework for efficiently evaluating SPARQL queries in a federated environment*". Federated query processing has been first illustrated through a running example in which a SPARQL query is launched over two distributed and heterogeneous data sources exposed as SPARQL endpoints, DBPedia, and The New York Times.

FedX has been introduced to address read-only scenarios through SPARQL 1.1 queries. No a-priori knowledge is needed to efficiently query multiple distributed data

sources (SPARQL 1.1 endpoints). Two main challenges have been identified to reduce the computational cost of federated query processing: the selection of relevants data sources and the computation of joins close to the data.
These challenges are tackled through a set of optimization techniques. *Source Selection* is realized through on-demand SPARQL ASK queries, caching information about the capabilities of each data source, in terms of triple patterns. *Exclusive Groups* consists in grouping together triple patterns relevant for a single data source. This allows delegating joins to the endpoints and thus, it limits the necessary network communications in the case of distributed joins. *Joins Reordering* is then realized based on a count heuristic. Finally, *Bind Joins* allow optimizing distributed joins through "vector" evaluations by using UNION (SPARQL 1.0) or VALUES (SPARQL 1.1) clauses. Experiments based on the FedBench benchmark show the efficiency of the approach compared to AliBaba and DARQ engines.

ANAPSID. Maria-Esther Vidal, from Simón Bolívar University (Caracas, Venezuela), discussed *On the Efficiency and Effectiveness of Federated Semantic Data Management*, and presented ANAPSID, an adaptive approach for federated semantic data management.
A first experiment shows that for a given set of distributed data sources and SPARQL queries, very different behaviours are observed. This motivates an in-depth analysis of queries and engines (e.g. FedX, SPLENDID, ANAPSID) to understand their impact on the performance of data federations. Data fragmentation is then introduced through horizontal fragmentation (each fragment potentially contains triples of many predicates) and vertical fragmentation (each fragment contains all triples of at least one predicate). Depending on the fragmentation scenario and the querying strategy, fragmentation may impact performance or results completeness. It has also to be noted that network latency has an impact on the performance of semantic data federations.
Two challenges for federated semantic data management are then highlighted : (i) formalizing query decomposition to better understand the characteristics of query engines and (ii) proposing optimization techniques to adapt to the dynamicity of the data sources (e.g. not being blocked if a data source becomes unavailable).
The ANAPSID engine tackles these challenges through a query decomposer and an adaptive query processor. Query decomposition is addressed as a vertex coloring problem. A DSATUR-based algorithm allows to group together triple patterns into subqueries based on joins and SPARQL endpoint capabilities. At runtime, the engine adapts it query plan based on the query operator behaviours and data source availability.
Finally, some limitations of existing benchmarks (e.g. FedBench) have been highlighted, especially in terms of query complexity, and network latency.

SemLAV. Pascal Molli, from University of Nantes (France), presented *SemLAV, Local-As-View mediation for SPARQL queries*. This work addresses data integration issues at the boundaries of the Linked Open Data space (matured RDF data) and the Deep Web, where data, even if open, originates from legacy and heterogeneous sources.
Three main approaches generally target data integration, namely *Wareousing*, *Global-As-View* (GAV), or *Local-As-View* (LAV). LAV approaches adapt to freshness and dynamicity requirements of the Semantic Web. They mediate data source heterogeneity through the generation of query rewritings. However, these query rewritings are difficult to handle in terms of the size of possible combinations. The main issue of LAV approaches is the explosion of possible rewritings, especially in the case of SPARQL queries that (i) possibly involve general predicates present in almost all data sources, and (ii) possibly involve star-shaped patterns.
The SemLAV approach consists in (i) selecting and ordering relevant views (i.e. having the maximum coverage), and (ii) materializing their content to allow SPARQL

queries to produce results incrementally. A benchmarking experiment (Berlin Benchmark, and a predefined set of rewritings) shows that, although the materialization leads to an increase of memory consumption, SemLAV produces more answers in the same amount of time, compared to state-of-the-art approaches.

AGGREGO Server. François Paulus, from the SemSoft company (France), presented and demonstrated the Aggrego platform. Through a *Local-As-View* approach, Aggrego addresses new challenges of data integration in the context of business/data intelligence. Enterprise legacy applications recently evolved to benefit from external data sources (social networks, open data, etc.) and they face nowadays heterogeneity, volatility or volume challenges.

Aggrego is proposed as a virtual database reconciling heterogeneous data sources. Aggrego relies on a global data model acting as a unified view over the heterogeneous data sources. Access to data sources is then dynamically orchestrated and relevant data is aggregated and homogeneously represented through the global data model. This resulting homogeneous data view is finally used to perform business queries, to populate databases or to run analytics tools. Aggrego's main use cases cover customer intelligence, social-media application and enterprise data integration.

Future directions. Works shown in this session addressed some of the federated data querying challenges, identifying several critical issues including:

- Distributed query performance optimisation;
- Reliability issue in a distributed setting where some of the sources might temporarily not be available; and
- Dynamicity of the federated data sources.

Performance optimisation of distributed queries is a complex and multi-factor problem in which sub-query evaluation plan and communication overhead play an important role. For instance, statistical approaches are being studied in FedX to improve data source selection and join re-ordering.

Furthermore, as-soon-as-possible delivery of results is often important to meet user expectations, as query systems are often expected to reply within seconds. To adapt to possible data sources unavailability, or first-n results scenarios, ANAPSID or SemLAV provide query answers progressively, as soon as they are available. Reliability of remote data sources access can be dealt with either through partial results downgrading or through data stores replication.

Integrating more diverse data source (sensor networks, updatable datasets) appears also as a challenging future direction (FedX, ANAPSID) in terms of data dynamicity.

6 Fifth session: Graphs and reasoning

Marie-Laure Mugnier and Jacopo Urbani presented their works to take into account ontological knowledge while querying data in graph models like RDF. They both considered reasoning with rules, either in backward or in forward chaining. Olivier Curé presented a complementary work on the storage of large-scale RDF triple stores handling inferences.

Marie-Laure Mugnier highlighted that the need for an ontological layer on top of data, associated with advanced reasoning mechanisms able to exploit the semantics encoded in ontologies, has been acknowledged in the database, knowledge representation and Semantic Web communities. She focused on the ontology-based query answering problem, which consists of querying data while taking ontological knowledge into account. To tackle this problem, she considers a logical framework based on existential rules, also called Datalog+/-. This emerging framework can also be defined as a graph-based framework. In her talk, she presented this framework and briefly reviewed the main decidability and complexity results, as well as algorithmic techniques, and pointed out two challenging research problems: querying over inconsistent data and Web-scale reasoning.

Jacopo Urbani focused on Web-scale reasoning. He highlighted that reasoning is a problem of primary importance in the Web, but also computationally expensive and thus hard to apply over very large amounts of data. He presented two approaches that demonstrated reasoning over the largest available inputs. WebPIE consists of a forward-chain reasoner based on MapReduce, and QueryPIE is a backward-chain distributed reasoner. Both approaches have demonstrated reasoning over inputs of 100 billion triples, which is roughly two or three times the size of the entire semantic Web. In his talk, he described the key principles behind the performance of these methods, and three lessons learned. (1) In his two approaches, the schemas are separated from the data and treated apart, replicated and kept in memory. (2) The data is range-partitioned, which is good for small queries (but bad for large queries). (3) Web-scale reasoning requires high engineering efforts, the choices of programming languages, libraries or compression techniques are crucial questions to reach certain performances; proof-of-concepts are not representative.

Olivier Curé highlighted the fact that RDF is a logical model that needs an appropriate scalable physical storage solution convenient for handling inferences related to some entailment regimes. He introduced the main strategies for storing and indexing RDF data sets mainly consisting in (1) solutions based on a native RDF approach, (2) solutions using a relational storage backend and (3) solutions using a NoSQL storage backend. Finally, he presented the main features of an original solution that aims to distribute highly compressed structures adapted for the storage and querying of RDF triples. In particular, its encoding of dictionaries supports the RDFS entailment regime.

In the framework of the CrEDIBLE project, we mainly focus on mediating and querying distributed heterogeneous RDF data sources. In this context, inferences can be limited to handling the semantics of RDFS vocabularies and these three presentations described solutions to handle RDFS entailment on big RDF data at query-time.
A secondary focus is emerging in the CrEDIBLE project: the secondary use of medical data once RDF datasets are aligned and linked into a single RDF graph. In this context, further inferences might be of interest to reason on these data, discover and explain new knowledge: OWL entailment and domain specific inference rules. This would lead to the general and more complex problem of handling inference rules over a large data set, addressed in the first two presentations. A main difference would be that in a secondary use scenario, inferences do not necessary need to be at query-time.