Mission pour l'Interdisciplinarité du CNRS MASTODONS
Défi Grandes Masses de Données Scientifiques

Projet **CrEDIBLE**
ConnaissancEs Distribuées en Imagerie BiomédicaLE

Document ID. CrEDIBLE-12-1-v1
Novembre 2012

CrEDIBLE Multi-disciplinary workshop

Sophia Antipolis, 15-17 October 2012

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

Summary

This document summarises the output of the CrEDIBLE multi-disciplinary workshop organized in Sophia Antipolis in October 15-17, 2012. 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 15-17 in Sophia Antipolis (France) where experts were invited to present their latest work related to biomedical data management and discuss their approaches. 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 4 thematic sessions over the 3 workshop days:

- **Data integration.** The data sources to be integrated are related but yet heterogeneous, using different semantic references (vocabularies…), different representations (files, relational / triple / XML databases…) and even different data models (relational, knowledge graphs…). Data integration will also be constrained by medical application constraints, in particular the set up of multi-centric studies, the support of translational research and medical applications. Data security and fine-grained access control is another important related problem. The ability to simultaneously process different data representation model makes data security a particularly challenging problem.
- **Ontologies.** The ontology defines conceptual primitives, which represent data semantics (images, test and questionnaire results) by integrating their production context (study, examination, subject, medical practitioner, data acquisition protocol, processing, acquisition device, parameterization, scientific publications). Such an ontology spans over different domains (different entity classes) and includes hundreds of concepts. It is structured through modules with different abstraction levels, to leverage generic primitives that can formalize several domains for practical reason related to ontology maintenance.

  The ontology design involves: reusing (completely or partially) existing ontological modules (at different abstraction levels); designing new modules (in particular to represent knowledge related to particular medical domains); managing modules life cycle; documenting to ease reusability. There are different means of exploitation: ontological alignment to federate data that rely on different semantics (addressing problems related to the level of details or even discrepancies in the entities considered); data processing assistance (checking the compatibility of data with processing tools, producing data provenance information); query-based and/or visualization-based data access. Each usage scenario might involve adapting the ontology representation to the tool manipulated (inference engine, visualizer) and its language.
- **Data representation models and reasoning.** Usually, medical data are stored in relational databases which allow for a fast access to data, while metadata are formalized through graph-based knowledge representation models, designed for the semantic Web, which enable reasoning capabilities through inferences based on ontologies used to model this knowledge. Main challenges are the mixed use of different representation and the scalability of data storage and reasoners. The scalability problem is well known in the Web of data community. Promising approaches lie on the use of graph-oriented databases, the adaptation of inferences performed to the size of manipulated

data stores, and on querying and reasoning techniques adapted to distributed stores.

- **Semantic workflows.** The acquisition and the representation of knowledge related to the manipulated data is tightly linked to the data processing and transformation tools applied. Knowledge acquired on data may be used to validate or filter the processing tools applied on this data. Conversely, knowledge acquired on processing tools can be used to infer new knowledge on data, in particular the data produced through this processing. Knowledge exploitation can happen at different levels of the scientific processing pipelines life cycle: at design time through editing assistance (static validation, assisted composition) and at run time (dynamic validation, new knowledge creation).

  Knowledge on both data and processing tools is also often used to describe data provenance information. Provenance is then described as semantic annotations tracing the execution path. Provenance is tightly related to the nature of data processed. It facilitates the reuse and the interconnection between data from different sources. It can make use of several domain ontologies and facilitate interoperability between different data processing engines.

1.2 Programme

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

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

2 Session 1: data integration

The first workshop session addressed the problem of medical data stores integration methods and tooling availability. The six talks presented covered the motivations for medical repositories integration, the challenges arising, and some methods that have been or are experimented in different contexts. Four operational platforms have been described (NeuroLOG, EHR4CR, I2B2, RICORDO). Other talks focused more specifically on methodologies or generic tools developed to achieve heterogeneous data stores mediation and federation.

The trend to integrate multiple clinical resources arises from the necessity to assemble data sets describing ever-larger patient cohorts in modern biomedical studies (e.g. tens of thousands patients needed in diabetes studies) and the difficulty to enrol patients in clinical trials. Cross-health enterprise studies design lead to challenging obstacles such as heterogeneity of data sources, data privacy control and performance of distributed data manipulation engines though.

2.1 Medical data integration examples

The NeuroLOG platform was developed to facilitate the sharing of neuroinformatics resources in multi-centric studies. It targeted Multiple Sclerosis, brain strokes, brain tumours and Alzheimer's in particular. To enable multi-centres data integration, it designed a data federation middleware that adapts non-invasively to heterogeneous legacy data repositories previously set up in neuroscience centres. This middleware is based on a domain Ontology (OntoNeuroLOG) that covers image acquisition, image representation, clinical tests and medical studies concepts. This ontology serves as a semantic reference for the federation. It is derived as a relational data schema for that purpose. A relational data federation engine (DataFederator from BusinessObjects/SAP) was then used to mediate legacy data sources and enable cross-federation relational querying (using SQL). In addition, the NeuroLOG initiative explored the ability to transform relational data stores into semantic repositories to enable semantic querying (using SPARQL).

Similarly, the EHR4CR project (Electronic Health Records for Clinical Research) targets the integration of clinical data (both Electronic Health Records and Clinical Data Warehouses), focusing on clinical trials design and execution life cycle. The middleware developed leverages existing standards (e.g. HL7 and CDISC) and domain terminologies (covering clinical findings, test results, laboratories, studies, medications...) to implement semantic interoperability between heterogeneous medical databases. A model-driven engineering approach was adopted, based on UML models to define a library of agreed data structure definitions, integrating various resources such as terminologies, ontologies, and other information models. Data queries are expressed using OCL and transformed into distributed SQL queries through pre-defined mappings, taking into account the terminology integrated in the platform. Query results are transformed back in a user-readable format through reverse mapping. The use of the SPARQL query language is considered for the future.

The I2B2 platform (Informatics for Integrating Biology and the Bedside) supports translational research in clinical genomics. The objective is to reduce clinical research costs by performing clinical trials as much as possible in-silico. I2B2 set up a centralized warehouse integrating more than 6 million patients and 1.5 billion anonimized diagnoses records. A simple star-shaped relational data schema cantered on the Patient concept is used to structure data. A client software

(Research Patient Data Registry) installed in healthcare centres can connect to both local and remote resources to select medical records of interest for each study. Once Institutional Review Boards have approved a study, selected medical records can be extracted to continue investigation. Specific protection measures against patient re-identification are implemented through fuzzy records extraction. Natural language processing queries are possible. Distributed resources assembly and querying are possible through an aggregator (SHRINE). In total, thousands of users registered to the I2B2 platform.

The Virtual Physiological Human RICORDO platform is another example of pharmacological data integration. It enables clinical data stores mining, exploiting domain knowledge and semantic technologies to align heterogeneous data sources. RICORDO developed a reference ontology through which source database entities can be annotated, thus making similarities between heterogeneous records explicit. In addition, the use of knowledge management technologies enables reasoning while querying.

In all platforms described, most data sources are relational databases. A strong emphasis is put on the semantic description of raw data records though. The NeuroLOG and the RICORDO platform abstracted data sources structure through semantic Web technologies to enable data alignment and querying (ontology-based data representation, mapping of relational entities to RDF triples, use of the SPARQL query language), while the EHR4CR platform adopted UML modeling, transformation and the OCL query language. The I2B2 approach favours data integration in a pre-defined data schema. Data distribution and distributed queries is considered in all cases. The sensitive nature of data makes it necessary to restrict the access to many data sources. It led to elaborated data protection mechanisms in the context of the I2B2 initiative.

2.2 Data mediation and querying techniques

Two approaches to semantic data stores distribution and querying were presented. SPARQL-DQP (Distributed Query Processor) and KGRAM (Knowledge Graph Abstract Machine) both enable the manipulation of distributed, heterogeneous databases and their querying through the SPARQL query language.

SPARQL-DQP is integrated in the OGSA-DAI (Open Grid Service Architecture - Data Access Integration) framework. It is completed with a relational-to-RDF mapping tool (based on the R2RML standard currently being finalized in a W3C working group) and an SQL query generator from SPARQL. It specifically addresses the optimization of distributed SPARQL queries through a query analysis and rewriting engine. It aims at providing more flexibility than the SPARQL 1.1 SERVICE clause to identify data records distributed over multiple databases without performance drop. In addition, SPARQL-DQP supports streamed data dynamic integration and querying.

KGRAM exhibits close functionality. It abstracts both the knowledge graph representation model and the query language to provide a versatile and flexible query engine. Semantic databases querying is implemented through graph matching, thus enabling SPARQL 1.1 compliant querying and other graph-based query languages (*e.g.* conceptual graphs). It also supports several entailment regimes and reasoning. KGRAM design features a modular software architecture, which facilitates the deployment of customized query processors adapted to various data manipulation scenarios. In particular, KGRAM supports multiple and distributed data sources querying through adapted components. Work to optimize queries in a distributed environment is on going. KGRAM also introduces an SQL query embedding mechanism in SPARQL queries, thus enabling the seamless integration of both relational and semantic data sources.

3 Session 2: ontologies, semantic modeling

The main goal of this session was to put in perspective general issues regarding the creation of ontologies and their reuse in specific applications. An important question is to assess the added value of foundational ontologies for both the design of new domain ontologies and the assembling of disparate ontologies into consistent application ontologies, dedicated to the needs of a specific application domain. The session allowed to consider this from several points of view: as foundational ontologies designer, biologist, clinical researcher and neuroimaging applications developer.

Claudio Masolo, from the Laboratory of Applied Ontology in Trento (Italy), presented the DOLCE-CORE foundational ontology and focused on the representation of properties. This aspect is very critical with regards to the applications of the biology / medicine sector since a major concern of this field is to characterize biological objects whose properties are observed and measured at multiple scales, at multiple time points and using various techniques. Precision in such a characterization is really critical with regards to relevant experimental data interpretation and reuse. Claudio situated the choices made in DOLCE-CORE with respect to different philosophical theories of properties, namely universalism, and trope theory.

The talk given by George Gkoutos (from the University of Cambridge) illustrated the use and the added value of ontologies for drug repositioning and suggestion of new drugs. He recalled basic models relating Genes, Gene functions, Involved pathways, Phenotypes, Actual physiological processes, Drugs and Diseases, a key feature for understanding the molecular basis of human disease, as well as the mechanisms of actions of drugs. In this respect George insisted on the need of matching pathobiology across different species and levels of granularity (*i.e.* addressing various facets of phenotypes, such as biochemical, cellular, anatomical and behavioural aspects). All these aspects can be characterized in a uniform and consistent way, through modelling their qualities. George introduced PATO, the Phenotype And Trait Ontology, and the different layers involved. PATO provides a vocabulary for the qualities that can then be related to the entities in which those qualities inhere, thanks to the so-called EQ model. This corresponds to the conceptual and semantic components layers that are then completed by a unification and integration layer to achieve the multi-domains and multi-species matching of experimental data. George illustrated how this allows predicting gene-disease associations and determination of gene functions, using results of the PhenomeNET. He explained the perspectives that this opens regarding novel drugs discovery and repurposing, especially with regards to future personalized interventional treatments.

The third talk was given by Jean Charlet (from Inserm UMRS 872 Eq 20 in Paris). The first part of his talk described works carried out in the context of the FP7 DebugIT project on antibiotic resistance. In this project ontologies are used to query, align and reason about heterogeneous data stored in data repositories located in several hospitals. Jean introduced the different ontologies involved, especially the DebugIT Core Ontology (DCO), 13 so-called operational ontologies (OO), supporting the mediation and integration layer, and the 7 data definition ontologies (DDO) used to model in RDF the relational schemas of the various databases. All these ontologies are consistently integrated using the BioTop ontology used as foundational ontology. The second part of the talk focused on the current re-engineering of the Orphanet classification system. The Orphanet database represents a classification of rare diseases. A key feature of this re-engineering is a better representation of the continuum of phenomenas, and better representing the relationships with genes whose presence can be associated to disorders (predisposing genes).

The fourth talk was given by Pierre Grenon, from the European Bioinformatics Institute in Hinxton (near Cambridge, UK), and involved in the RICORDO project. Pierre's talk aimed at presenting the development of ontologies for the annotation of biomedical models and data. Pierre recalled first general aspects of ontology development, insisting on the importance of clearly defining the goals and purposes. He then introduced an ontology for the models, allowing representing the entities composing the model (*e.g.* the compartments, or the parameters of a model and their physical interpretation), as well as the thematic domains involved, the computer programs implementing it. Pierre illustrated his talk with a few samples from the domain of pharmacokinetics.

The final talk was given by Bénédicte Batrancourt, from INSERM UMR_S975 at the Pitié Salpêtrière (ICM) in Paris. Bénédicte introduced the CATI project, a central resource set in Paris Pitié Salpêtrière and Saclay (CEA Neurospin) to support image management and processing of images in the context of the Alzheimer's disease. CATI is a service platform providing its services to a network of 25 clinical centres in France and complements the MEMENTO cohort. Bénédicte provided some details about the implementation of CATI, how CATI manages the various stages of the workflow: image upload, data de-identification, quality control, storage in the CATI Buffer, CATI Shared and CATI Cluster repositories. She went into more details regarding the database schema (CATISchema), articulated to an ontology called OntoCATI, based on the OntoNeuroLOG ontology as well as other sources such as DICOM and XCEDE.

4 Session 3: data representation model and reasoning

In the continuation of the first session, Rémy Choquet and Konstantin Todorov presented works on **ontology-based heterogeneous data or knowledge integration**. Rémy Choquet presented the approach used in the EU project DebugIT to make antibiotics resistance data semantically and geographically interoperable. It relies on a set of ontologies that helps integrating and comparing data from 7 European hospitals. Konstantin Todorov presented an approach to bring together heterogeneous data by bridging the gap between the vocabularies that describe them. It consists in building a common fuzzy ontology from a set of domain ontologies, based on *fuzzy sets*. Every domain concept is represented as a fuzzy set of the concepts of a particular reference ontology and a fuzzy subsumption relation is defined over these fuzzy concepts.

Marie-Aude Aufaure and Ched Raïssi addressed the problem of **extracting knowledge from large datasets**. Marie-Aude Aufaure addressed the problem of extracting and managing large knowledge graphs extracted from relational data. She presented an overview of graph databases and distributed computing. She introduced a set of solutions to extract graphs from structured data and facilitate the process of information search in these graphs as well as their aggregation for community detection and visualization. Finally, she depicted a solution for merging graphs using the Hadoop/Map Reduce framework. On the other hand, Ched Raïssi addressed the problem of the queries complexity. He introduced the notions behind the concept of Knowledge Discovery guided by Domain Knowledge and presented a pattern mining approach based on Formal Concepts Analysis (FCA). Starting from the notion of Skyline queries, he presented an approach to compute a compressed skycube storing all subspace skylines. This could be used to query the Semantic Web. Finally, He presented an approach to discover skyline patterns.

Pascal Molli and Sébastien Ferré finally addressed two problems related to **searching the Web of Linked Data**. Sébastien Ferré addressed the problem of finding an expressive but user adapted way of searching the web of linked data. He started from the assessment that querying languages, such as SPARQL, offer expressive means for searching RDF datasets but are difficult to use, while faceted search supports exploratory search but does not offer the same expressiveness as query languages. As a result, he presented an approach to reconcile expressive querying and exploratory search where the navigation of faceted search is formalized as a navigation graph, navigation places are queries, and navigation links are query transformations. Pascal Molli addressed the problem of making Linked Data writable: taking into account the continuous updates of some RDF stores when searching the web of linked data. To query data distributed over several RDF stores either requires to copy datasets locally or to perform distributed querying. Starting from the assessment that local copies have problems of freshness and distributed queries problems of scalability and performance, he proposed an approach to make RDF stores live by providing streams of data updates: each Linked Data node can follow update streams of others, creating a social network of live updates. This opens a third way to query by synchronizing and searching, with concurrency and consistency issues to be addressed.

5 Session 4: semantic workflows

The last session of the CrEDIBLE workshop addressed the use and the enrichment of knowledge repositories trough data processing pipelines (workflows). Freddy Lécué (IBM Research, Dublin, Ireland) first addressed the design of workflows by composing semantic web services. Paolo Missier (Newcastle University, UK) addressed the problem of generation and exploitation of provenance information resulting from workflow runs. Finally two preliminary works addressing information overload in e-science activities have been presented by Nadia Cerezo and Alban Gaignard (I3S, Sophia Antipolis, France).

Composing and Optimising Services in the Semantic Web

Freddy Lécué presented some contributions towards optimal web-based services composition, addressing challenging automation, dynamicity and scalability problems. Semantic web principles are applied to web services. Services are described at functional level in terms of *Input* and *Output* parameters (based on SAWSDL, OWL-S profiles, or WSMO description languages), and in terms of *Preconditions* and *Effects* through Horn-like rules (based on the SWRL rule language). Several levels of semantic connection (matchmaking functions) are considered between two services (*exact, plugin, subsume, intersection, disjoint*). The main issue appears when the matchmaking is imperfect. It requires negotiating or discovering the missing information to achieve the match. The automated service composition is achieved through a semantic augmented artificial intelligence planning approach (goal-based reasoning). It consists in reasoning on the domain ontology (the TBox of the knowledge base) to propose semantic links between service descriptions, and reasoning on semantic service descriptions (the ABox of the knowledge base) to validate instances. Finally industrial use cases and scalability experiments have been presented.

Workflows, experimental findings, and their provenance: towards semantically rich linked data and method sharing for collaborative science

This talk addressed the setup of collaborations in virtual experimental sciences. Properly reusing scientific data formerly generated in the context of a different study requires a comprehensive set of metadata describing the generated data semantics. Provenance traces, captured through the invocation of scientific workflows, form a basis for a better understanding of the data processing history. The Janus semantic provenance model extends the PROVENIR domain-agnostic upper ontology with domain-specific concepts involved in bioinformatics Taverna workflows. As soon as inputs and output ports of workflow processors are annotated with domain concepts, the system is able to propagate the semantic annotations to the produced data. Finally, produced annotations are mapped to Linked Open Data through the generation of URIs incorporating Bio2RDF Linked Open Data sources.

The PROV-O provenance ontology standardized by W3C was introduced. PROV-O core elements are *Activity*, *Entity* and *Agent* and their relationships. A specificity of PROV-O is its ability to deal with nary relations through reified relationships.

The last part of the presentation was dedicated to a short introduction on *Research Objects* addressing data, methods, and provenance packaging and sharing in the context of the Wf4Ever and the DataOne projects.

VIP semantic workflows

The last presentation of this workflow session addressed information overload in e-science activities through two preliminary works. Scientific workflows design and exploitation require expert skills. Because of several intertwined granularities and representation layers, and several end-user activities, it is challenging to provide relevant information to all e-science platform users. The general objective of these works is to help e-scientists focussing on meaningful information by (i) exploiting data (and associated domain knowledge) in e-science experiments, and (ii) exploiting e-science workflows (and associated domain knowledge).

Provenance information from workflow runs can help e-scientists in determining the cause of failure or abnormalities and propagate knowledge on data and processing tools to the produced data. The final objective is to produce meaningful experiment summaries. These summaries result from domain-specific inference rules taking as input both semantic service descriptions and fine-grained provenance information (using the OPM provenance model). The produced statements form a meaningful experiment summary involving few but meaningful domain-specific statements associating domain concepts and properties to the produced data.

The design of scientific workflows can be simplified through a goal-based approach. The objective is to generate workflow descriptions from conceptual descriptions based on high-level workflow *Fragments*. A Fragment is composed by a *Pattern* defining the fragment weaving points, and a *Blueprint*, describing a flow of activities to be injected into the target workflow. Conceptual workflows are derived into concrete workflows by weaving blueprints.

The last part of the talk integrated these two approaches as perspectives. Conceptual workflows could help in the design of provenance-based inference rules and enhance their genericity. Conversely, provenance information could be used to suggest annotations at conceptual workflow design-time based on produced and annotated data.