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-14-2-v1
Décembre 2014

CrEDIBLE Multi-disciplinary workshop

Sophia Antipolis, 8-10 October 2014

C. Faron Zucker, A. Gaignard, B. Gibaud, F. Michel, J. Montagnat

Summary

This document summarises the output of the third CrEDIBLE multi-disciplinary workshop organized in Sophia Antipolis in October 8-14, 2014. This 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 open scientific questions.

Table of content

1 WORKSHOP OBJECTIVES 3

1.1 SUMMARY 3

1.2 PROGRAMME 3

2 FIRST SESSION: BIOMEDICAL DATA FEDERATION IN PRACTICE: FEEDBACK ON EXISTING APPROACHES 4

3 SECOND SESSION: REASONING AND VISUALIZING LARGE DATA GRAPHS 6

4 THIRD SESSION: DATA FEDERATION 7

5 FOURTH SESSION: DATA MEDIATION 8

6 FIFTH SESSION: BIOMEDICAL ONTOLOGIES 11

1 Workshop objectives

The CrEDIBLE project organized 3 multi-disciplinary working days in October 2-4, 2014 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. A moderator from the CrEDIBLE consortium led each panel discussion.

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

- **Biomedical data federation in practice: feedback on existing approaches.** This session reports on concrete experience in developing systems gathering or indexing data to be shared and reused in the context of research projects. It discusses user requirements, success stories, current technology limitations and future expectations.
- **Reasoning and visualizing large data graphs.** Once data from heterogeneous and distributed stores has been mediated and federated, exploiting the resulting data implies, among others, to visualize and reason over large graphs. There is a trade-off to be found between the amount of data to process and the reasoning capabilities of the system.
- **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.
- **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 kind of data are considered.
- **Biomedical ontologies.** This session discussed ontologies modelling of medical data interoperability through ontological representations. Different modelling approaches, both top-down and bottom-up where reported.

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, Biomedical data federation in practice

- **Camille Maumet** (University of Warwick, UK): Supporting image-based meta-analysis with NIDM: Standardised reporting of neuroimaging results

- **Silvia Olabarriaga** (Amsterdam Medical Center, NL): Challenges for services integration into science gateways: the local story of the AMC
- **Pascal Neveu** (INRA Montpellier, FR): From Gene To The Bottle
- **Frédérique Segond** (Viseo, FR): Language and knowledge technologies to properly model textual medical data and support better reasoning

Session 2: Reasoning and visualizing large data graphs

- **Harald Sack** (U. Potsdam, DE): The Journey is the Reward - Exploratory Semantic Search based on Linked Data
- **Martin Peters** (FH Dortmund U., DE): Rule-based reasoning using GPUs
- **Guy Melançon** (U. Bordeaux): Visual Analytics Supporting Rule Based Modeling
- **Christian De Sainte Marie** (IBM): Business rules engines

Session 3: Data federation

- **Ruben Verborgh** (Ghent University): Querying data on the Web – client or server?
- **Ester Pacitti** (U. Montpellier 2 / INRIA, FR): Profile Diversity for Query Processing using Users Recommendations
- **Axel-Cyrille Ngonga Ngomo** (U. Leipzig): HIBISCUS: hypergraph-based sources selection for federated SPARQL queries

Session 4: Data mediation

- **Pascal Molli** (U. Nantes, FR): Towards Writable and Scalable Linked Open Data
- **Franck Michel** (CNRS, FR): xR2RML, an R2RML extension for the translation of non-relational databases into RDF
- **Freddy Priyatna** (U. Polytechnic Madrid, SP): An Overview of the Research Carried Out at Data Integration Group - OEG

Session 5: Biomedical ontologies

- **Simon Jupp** (EBI, UK): Interoperability of large-scale image data sets with ontologies
- **Heiner Oberkampf** (Siemens, DE): Expressing Medical Image Measurements using Open Biological and Biomedical Ontologies
- **Alan Ruttenberg** (University at Buffalo, School of Dental Medicine, USA): What should the role of biomedical ontology in imaging be?
- **Bernard Gibaud** (INSERM, FR): Ontology modelling needs for image biomarkers description

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: Biomedical data federation in practice: feedback on existing approaches

This session gathered four talks delivering feedback on different data integration use cases. The two first talks, by Camille Maumet (University of Warwick) and Silvia Olabarriaga (Amsterdam Medical Center) both focused on medical image databases, in the context of neuroimage analysis pipelines and a broad scope experiment support portal respectively. The third talk, by Pascal Neveu (INRA Montpellier), targeted wine plantations data federation through the use of a coherent ontology set. Finally, the fourth talk, by Frédérique Segond (Viseo), addressed free text medical records analysis and exploitation.

In her talk entitled “Supporting image-based meta-analysis with NIDM: Standardised reporting of neuroimaging results”, Camille Maumet summarized an activity conducted in the context of the INCF NeuroImaging task force which aims at developing standards and tools for neuroimaging data and associated metadata sharing among different applications. This group designs a NeuroImaging Data Model (NIDM), inspired from the BIRN XCEDE data model, that is an extension of the W3C PROV standard in the domain of human brain mapping. Neuroimaging data provenance information is captured at every steps of neurodata manipulation: acquisition, analysis, post-processing, publication and archiving. This rich metadata set can be used to increase data reuse opportunities and statistical power of data analysis pipelines. Meta-analysis can be assembled by considering results obtained from different data sets acquired in the context of different experiments. Furthermore, there is an increasing trend to share among scientists not only the (very small) set of results summarized in publications, but also raw images and detailed image-based analysis results. The data sharing environment and tools developed in this context are integrated in the three most used software packages in neuroimaging: SPM, FSL and AFNI (with different implementation status). Pipelines written using different software packages can thus be described with a common terminology. Work continues for completing the implementation in the AFNI software package and to integrate more vocabularies (eg. BirnLex).

Silvia Olabarriaga presented the experience gained in developing several generation virtual research environment to support AMC researcher needs, in her talk entitled “Challenges for services integration into science gateways: the local story of the AMC”. The COMMIT e-Biobanking project[1] reported many success stories in the manipulation of semantic data for medical applications as different as data provenance tracing, visualization of many drugs/proteins possible interactions, data reuse and large-scale data analysis. The development of science gateways at AMC, integrating as much functionality as possible to match user expectations, has been a long run behind changing technologies and computing environments over the past decade though. It resulted in 5 generations of science gateways from the most basic grid access environment to today’s web portal with support for multiple computing infrastructures, different data sources and semantic-rich metadata. Important milestones have been the integration of provenance information in the platform, interface with the XNAT data model and performance bottlenecks removal. Many open issues remain though as some trade-off are not easily found to build a balanced platform (e.g. granularity of provenance versus amount of provenance information; archiving of data versus re-computation; vocabularies to use among the existing ones…).

Pascal Neveu, from the French National Agronomy Research Institute (INRA), gave the next speech entitled “From genes to bottle”. It addressed the problem of data collection and integration in the context of wine production, which is a representative example of a complex food transformation process. The data collected is used in climate change and production impact studies. This data is heterogeneous describing information as diverse in domain and scale as genetic material, weather conditions, farming practices and grapes fermentation. Furthermore, the data acquisition processes set up in different contexts are usually not designed to facilitate data sharing and rarely cross-over discipline boundaries. The objective of this work was to reduce the gap between disciplines and automate knowledge collection from heterogeneous data sources leveraging semantic data description technologies. A set of ontologies was developed (e.g. for viticulture and wine transformation) based on upper-level ontologies (DOLCE, SUMO) and existing ones (e.g. W3C Time Ontology, Agrovoc, Plant Ontology…) to facilitate reuse. Multi-devices domain tools were developed to import data from various sources into RDF graphs, visualize it, cure it and complete it. They make it possible to assemble collaborative knowledge graph capitalizing on information collected in the entire wine making process. It is expected that this approach can generalised to other plantation studies.

[1] E-Biobanking project: http://www.commit-nl.nl/projects/e-biobanking-with-imaging-for-healthcare

Frédérique Segond gave the last session talk entitled “Language and knowledge technologies to properly model textual medical data and support better reasoning” on electronic medical records (EMR). In spite of the emerging information society, most medical reports remain unstructured today and there is a growing need to transform textual data into EMRs automatically for better data exploration and data reuse in epidemiological studies. Several problems related to free text-to-EMR data transformation have been investigated in the context of the Synodos project[2], such as (i) the conceptual data model to use (which level of details need to be tuned); (ii) the clear definition of an episode of care (which is usually spread over several reports); (iii) the categorization of medical terms (to differentiate terms relating to patient history from terms relating to symptoms or diagnosis for example); and (iv) the identification of frontiers between disciplines. The free-text reports processing starts with a linguistic analyzer specialized in the medical field to generate syntactic dependency graphs bound to the medical terminology. More facts are then deduced by applying transition rules (gender or temporality between events can be inferred directly from textual information…) and template-based expert rules designed by medical experts (surgical infection can be inferred from post-operative treatment…). The problems of querying such a database and making results understandable for end users remains a challenging issue. There are interesting perspectives in integrating other kind of data such as image-based and biology sample-based data.

The session was concluded by a panel discussion driven by questions on metadata collection, agreement on standards and data provenance capture. Metadata collection is a problem that often requires the collaboration of end-users. The value of annotating raw data with metadata is not always sufficiently recognized to ensure that data providers are willing to spend the time needed to enrich the data. Existing standards and appropriate tools definitely help in collecting the data but even when they are available, there might be very different views on important metadata in different sub-domains. The promotion of standards is an important factor but it is often difficult to identify the standards likely to be widely adopted. In the medical context, some standards such as SNOMED are the propriety of a company and standards proposed by large organization are not necessarily the better-adopted ones (some widely adopted tools create de-facto standards). Finally, provenance is an important source of metadata. The capture of provenance traces raises non-trivial questions on the granularity of traces to collect (as provenance traces may sum up to large amounts, even compared to raw data) and planed usage for this information.

3 Second session: Reasoning and visualizing large data graphs

After mediating and federating data from heterogeneous and distributed store, the question arising is the exploitation of the large data graph available. This implies, among others, to enable reasoning over it and visualizing it. This is a keystone to translational research.

While the Web of data focuses on large data sets processing, the semantic Web involves costly reasoning processes. There is a trade-off to be found between the amount of data to process and the reasoning capabilities of the system. Meanwhile, information visualization aims at providing methods and tools to access and explore large amounts of data, at conveying knowledge from these data in intuitive ways and therefore understanding them. For the Web of data, we focus on the visualization of large data graphs. As a result, the research questions addressed in this session were the following:

[2] Synodos project: http://www.viseo.com/en/synodos-project

- How to visualize and navigate on large data graphs for exploratory search, visual mining and understanding of these data?
- How to apply inference rules on large distributed data graphs?
- How to reconcile visualization of data and automatic reasoning?

The presentation of Harald Sack from the University of Potsdam (Germany) addressed the first question. He presented an approach and a tool, Yovisto, dedicated to exploratory search based on DBPedia. He argued that, if the user is not familiar with the search domain or if the information need of the user requires subsequent queries that build on one another, traditional retrieval technology reaches its limits. Exploratory search exploits the content-based relationship of documents to enable the discovery of relevant results along the guided search path. Moreover, it enables also serendipitous discovery of solutions to a problem that is relevant but not intentionally thought of. This approach could be adapted to translational research, based on an exploratory search among an integrated RDF database.

The presentation of Martin Peters from the FH University of Dortmund (Germany) addressed the second question. He presented an approach for implementing a rule-based reasoner that uses the massively parallel hardware of modern graphic cards (GPUs) to reason on large RDF graphs. He showed how the RETE algorithm, which is a pattern-matching algorithm that can be used to implement production systems, can be adapted for a highly parallel execution on the GPU. Based on the introduced concepts the materialization of the complete RDFS closure for 1 billion triples can be performed on a single computing node, reaching a throughput that is comparable to state of the art MapReduce-based approaches.

The presentation of Guy Melançon from the University of Bordeaux (France), addressed the third research question. He presented his work on building a visual analytics system supporting rule-based modelling. The PORGY framework built and distributed in his team supports tasks that were carefully identified as central to modelling with rules. It aims at designing relevant graphical representations and adequate interactions on dynamic graphs emerging from graph rewriting systems.

The presentation of Christian De Sainte Marie from IBM France Lab addressed the second and third questions. He first presented the JRule Business Rule Management System developed at ILOG, for both collaborative rule management and rule execution. Then he presented the Ontorule research project, aiming at combining ontologies and rules in IT applications and addressing the problem of the acquisition of ontologies and rules from the most appropriate sources, their separate management and maintenance; and their transparent operationalisation.

4 Third session: Data federation

Profile diversity for user recommendations

Esther Pacitti, from the Zenith research unit (University of Montpellier 2, INRIA, CNRS) discussed the exploitation of the user profile diversity, in the context of citizen sciences, to enhance searching and recommendation of scientific artifacts. Relevance of recommended items depends on the diversity of their content. As an example, while a few plants represent the majority of observations, the majority of plants are rarely observed. The objective of this work is to find a good trade-off between relevancy and diversity. The approach proposes a gossip distributed protocol exploiting user profile diversity, which is implemented in the context of the Pl@ntnet platform.

Linked Data Fragments

Ruben Verborgh, from the Ghent university (Belgium), presented a talk on “Querying data on the Web: client or server?”. The presentation first illustrated the main issues faced when querying publicly available linked data sources. Nowadays, public SPARQL endpoints offer a rich querying interface. In some cases, complex queries cause their overloading and as a consequence, it strongly impacts the availability and reliability of SPARQL endpoints. The *Linked Data Fragments* approach has been introduced to address the scalability and reliability of semantic web data querying, by adjusting the balance between client and server efforts. *Linked Data Fragments* are specified through a *selector* to define content in terms of triples, some *metadata* describing the fragment, and some *control statements* to access the other data fragments. Query efficiency comes from selecting first the smallest fragment (count metadata). The proposed client-side SPARQL querying implies different trade-offs compared to classical semantic web infrastructures relying on few SPARQL endpoints. It comes with increased network bandwidth consumption and slower results, however it offers a better availability of servers through low-cost but simpler queries, and streamed results.

HIBISCUS

Axel-Cyrille Ngonga Ngomo, from the AKSW research group (Germany) presented HIBISCUS, “an hypergraph-based source selection for federated SPARQL queries”.

The presentation first introduced the SPARQL federation process through an illustrative example. It consists in parsing an input SPARQL query, selecting the relevant data sources, distributing with optimizations sub-queries over the distributed data sources, and finally integrating the results. A common issue in federated SPARQL querying is the overestimation of the number of contributing data sources, which leads to extra network traffic generated, as well as increased query execution time. To address this issue, HIBISCUS focuses on data source selection and aims at identifying capable and contributing data sources with regards to individual triple patterns. HIBISCUS uses the URI authorities (dbpedia.org in http://dbpedia.org/ontology/) and data summaries (representing the capabilities of data sources) to prune irrelevant data sources. HIBISCUS has been evaluated through the FedBench benchmark as an extension of the FedX, DARQ and SPLENDID state of the art federated SPARQL engines. Regarding query execution time, results show 92% of performance improvement compared to DARQ, 82% of performance improvement compared to SPLENDID, and 24% compared to FedX.

Summary

Works presented in this session mainly addressed the performance, the scalability and the reliability of linked data querying. Performance, in terms of query execution time, is tackled through the accurate selection of relevant data sources. Based on data source capabilities and URI authorities, HIBISCUS proposes source selection as an independent component extending state-of-the-art federation engines. To tackle the scalability challenges of distributed linked data querying, the Linked Data Fragments approach proposes to shift the querying load from the server to the client. The counter-part is slower queries and increased network communications, but the main advantage is a better availability of SPARQL endpoints and thus an enhanced reliability of linked data applications.

The panel discussion highlighted that FedBench is the only available benchmark for federated SPARQL querying, which is a good starting point. However, it is not representative of very large-scale applications such as the cancer genome atlas (TCGA), which points out the need for larger scale benchmarks.

5 Fourth session: Data mediation

Mediation refers to the ability to overcome the mismatch between formalized data stored in heterogeneous data sources, which were often designed independently from each other. This mismatch can result of the heterogeneity in terms of representation languages, terminologies used, model discrepancies, varying scopes and points of views. Resources being expressed in different ways (using different data models) must be reconciled, i.e. semantically aligned, before they can be used together. Data Mediation is often a pre-requisite for Distributed Querying.

Three talks were given during this session. The first one deals with ways to achieve a writable linked open data, while the second and third presented approaches meant to translate heterogeneous data sources into RDF.

Pascal Molly (GDD, University of Nantes) first gave a talk entitled “Towards Writable and Scalable Linked Open Data”. Linked Open Data (LOD) suffers from data quality and availability issues. For example, broken links occur when a data provider changes IRIs of resources referenced in other data sets. Involving data consumers can be a key to fixing such issues, in some kind of social organization for writing LOD. However, the classical open-data environments only make data accessible for reading and changes cannot be made. Pascal presented collaborative graph (Col-Graph) and FEDRA, two attempts to involve data consumers and improve data availability and quality.

The *Col-Graph Network* solution uses linked data fragments, i.e. partial copies of data sets. On the one hand, an incremental maintenance of fragments consumes change logs of sources to propagate updates. On the other hand, data consumer can update a fragment to fix data inconsistencies before pushing a patch proposal to the original source. A synchronization protocol ensures the eventual consistency of the Col-Graph network. The Col-Graph approach enables the creation of a read-write Linked Data that improves general linked data quality in a crowd-sourcing approach.

Existing federated query engines such as FedX and Anapsid hardly exploit data sets replication and fragmentation, although that knowledge can help addressing the source selection question. FEDRA takes advantage of fragments definition to deduce containments and perform source reduction at query-time. Overall, FEDRA can be used along with other federated query engines to improve scalability and availability of LOD. Some tests show that FEDRA's approach allows answering complex queries that FedX and Anapsid either fail to process or process in a time frame that make them unusable.

Franck Michel then presented **xR2RML**, an R2RML extension for the translation of non-relational databases to RDF. The web of data is progressively emerging along with the publication and interlinking of various open data sets in RDF. Its success largely depends on the accessibility of semi-structured documents and structured databases in a common format. In particular, NoSQL systems have gained a remarkable success during recent years but the data it contains mostly remains "locked". Other types of databases have been developed over time, such as XML databases (notably used in edition and digital humanities), object-oriented databases or directory-based databases. Significant efforts have been invested in the definition of methods to translate various kinds of data sources into RDF, such as R2RML, an RDB-to-RDF W3C recommendation, and RML, an extension of R2RML to address the mapping of various data formats, in particular hierarchical formats (XML, JSON). Thus, although heterogeneous data formats are addressed no common language describes, in a uniform manner, mappings of relational and non-relational databases to RDF so far.

xR2RML is a mapping language that extends R2RML and RML. It addresses the mapping of a large, extensible, scope of non-relational databases to RDF. It allows querying almost any database without any assumption as to which query language is being used. It relies on RML concepts to reference data elements within query results independently from the underlying data model (row-based in RDBs or BigTable-like NoSQL systems, hierarchical models such as JSON in NoSQL document databases). Additionally, it extends R2RML and RML capabilities with two features: (i) it can deal with contents of mixed nature, like for instance some JSON or XML data embedded in cells of a relational table; and (ii) it extend the scope of RDF terms generated with the ability to create RDF collections (rdf:List) or containers (rdf:Bag, rdf:Seq, rdf:Alt) from list of values. A prototype implementation of xR2RML is currently being developed, that addresses RDBs and one example of NoSQL document database, MongoDB. An open question concerns the management of graph data models: can a hierarchical data model satisfyingly approximate a graph?

The last talk by Freddy Priyatna gave an overview of research at the data integration group from Madrid Polytechnic University. Freddy presented Morph, a suite of technologies focused on the translation of heterogeneous data sources into RDF and their querying. Morph investigates query-rewriting techniques through the use of mappings expressed in the W3C R2RML language. The suite consists of the following tools:

- morph-RDB for accessing relational databases. It has been used to deal with various real-world biomedical queries.
- morph-LDP, an extension of morph-RDB that enables the read-write access to LDP4j, an Linked Data platform implemented at OEG. In particular, morph-LDP enables a SPARQL-based update to a LD platform backed by a relational database. To do so, it implements a SPARQL-Update to SQL query rewriting process using an R2RML mapping description.
- morph-streams, for accessing data streams produced by querying sensor, using SPARQL extension (SPARQL-stream). morph-streams++ is an ongoing work targeting a more scalable RDF stream processing engine by parallelizing the processing of query operators.
- SPARQL-DQP, a federated SPARQL engine.
- morph-GFT, for accessing Google Fusion Tables.
- kyrie, a reasoning engine for expanding SPARQL queries by considering ontology entailments (aka. ontology-based query rewriting).

Overall, this session addressed two main issues that are discussed below: the translation of “not-only relational data” to RDF, and the achievement of a writeable web of data.

With the data deluge going on ever faster, it is crucial to come up with methods that can expose databases in RDF. The work around relational databases is now mature with mapping languages like R2RML and implementations such as morph-RDB. However, there has been, during the last years, a trend to migrate from one-fits-all systems (RBDs) to multiple heterogeneous systems: XML databases are often used in digital humanities and edition; the manifold NoSQL databases, initially driven Big Data needs of major web players, have very different data models and query capabilities (key-value stores, document stores, extensible column-stores aka. BigTable-like systems, graph stores). xR2RML is an example of approaches dedicated to dealing with such a wide scope of database systems. More and more data sensors are being deployed in smart cities, smart houses, etc, continuously producing streamed data. Morph-stream is an example of such streamed data management.

Although those approaches succeed in publishing RDF data from heterogeneous databases, they most generally provide read-only data. And similarly to the web 2.0 that marked the transition from a static web to a web in which any user produces new data, the web of data will need to be write-enabled at some point. Two examples were presented in this session. Morph-LDP investigates the rewriting of SPARQL-Update queries to SQL queries, based on an R2RML mapping description. This solution applies in a well-controlled context: within a company for instance, with authenticated users and well know applications, authorized to do such updates. Yet, any kind of SPARQL update cannot be performed since an R2RML mapping is not always reversible. For instance, using SQL aggregate functions AVG, SUM...) in an R2RML mapping can be used to convert relational data into RDF triples, but reverting from RDF to the original data model is not possible.

In the more general Linked Open Data context, we must consider that SPARQL endpoints (public endpoints typically) are read-only. Thus, enabling a read/write web of data should consider other options. Col-Graph is one of them: it assumes that partial copies of data sets should be made here and there, representing fragments of data sets. Fragments are updated by data consumers and synchronized with original data sources by consuming logs. Besides, it is possible to take advantage of fragments to improve linked data availability and scalability by distributing multiple replicates of the same data. This is the approach proposed by FEDRA.

6 Fifth session: Biomedical ontologies

The fifth and last session was composed of four talks addressing different and complementary aspects of biomedical ontologies' design.

The first presentation entitled “Interoperability of large scale image data sets with ontologies” was given by Simon Jupp of EMBL-EBI, UK. Simon introduced the BiomedBridges project, a major FP7 project connecting most of the research infrastructures projects setup by the ESFRI (European Strategy Forum on Research Infrastructures) to boost and synergically develop major infrastructures in biology and life science research at the European scale. Simon first introduced the problem to be faced in BiomedBridges, namely ensuring consistent naming of biological features of interest throughout the many databases and the different scales at which biological phenomena may be described. He underlined the key role of imaging in articulating together such descriptions at a cellular, gene, and molecular levels, across different species (e.g. study at cell level in animal models and human cancer tissue). He explained that existing ontologies were not sufficient to address these issues and he explained why and how the Cellular Microscopy Phenotype Ontology (CMPO) was developed. This was clearly a bottom-up approach consisting of collecting phenotype-ontology mappings (based on GO and PATO) and converting such annotations to OWL axioms. He concluded by mentioning the setting up at EBI of a large RDF platform accessible as SPARQL endpoints and supporting all their databases.

The second presentation entitled “Expressing Medical Image Measurements using Open Biological and Biomedical Ontologies” was given by Heiner Oberkampf from Siemens, Germany. Heiner introduced first some typical examples of measurements met in radiology and pathology. He then explained in details how the Ontology of Biomedical Investigations (OBI) could be used to model such measurements, e.g. specify the quality being measured, the object bearing this quality, the value and unit of measure of the measurement, etc.

He presented extensions that he had proposed in his PhD work to model entities that are not present in OBI such as regions of interest as well as design patterns for expressing normal upper bound and normal interval specifications and relating them to reference populations.

The third presentation entitled “What should the role of biomedical ontology in imaging be” was presented by Alan Ruttenberg from the University of Buffalo in the

US. Alan presented in detail the Open Biomedical Ontologies (OBO) foundry. He listed the basic principles underlying the creation of a set of non-overlapping ontologies in the domain of life sciences. He presented in detail the Basic Formal Ontology (BFO), a foundational ontology designed in Barry Smith's group and used as a basis for all the OBO ontologies. He presented a number of illustrative applications in the domain of dentistry.

The fourth presentation entitled “Toward ontologies for imaging biomarkers description” was presented by Bernard Gibaud from Inserm LTSI, France. Bernard introduced the notion of imaging biomarker and explained the key role they are going to play in clinical and translational research, on the one hand, and in care delivery on the other hand. He then described the domain to be covered by such ontology of biomarkers, and detailed which existing ontologies could provide meaningful starting points for developing it. He reviewed a number of such ontologies (OBI/IAO, RadLex, OntoNeuroLOG, OME, QIBO) and highlighted which were the major gaps and challenges.

The session was followed by a discussion highlighting the very different approaches used in the works that were presented: bottom-up approach used in BiomedBridges with the goal of rapidly providing ontologies that can support the sharing of phenotypes descriptions, compared with the rather top-down approach used in OntoNeuroLOG with an effort to produce consistent ontological models valid for all kinds of medical image modalities, or even microscopy imaging. The former approach appeared quite close to the one currently used in the International Neuroinformatics Coordinating Facility (INCF) consortium for managing and sharing neuroimaging data (presented the day before by Camille Maumet), with strong emphasis on delivering rapid results can may be used in annotating image and statistical data produced by widely used software packages such as FREESURFER, FSL or SPM. The latter approach was closer to the OBO way to design ontologies, with a sound philosophical basis. The question is whether such modelling approaches can be reconciled and actually converge, which seems rather doubtful.