Interoperability driven integration of biomedical data sources

Douglas TEODORO[a,1], Rémy CHOQUET[c], Daniel SCHOBER[d], Giovanni MELS[e], Emilie PASCHE[a], Patrick RUCH[b], and Christian LOVIS[a]
[a]*SIMED, University Hospitals of Geneva and* [b]*HEG, University of Applied Sciences, Geneva, Switzerland;* [c]*INSERM, Université Pierre et Marie Curie, Paris, France;* [d]*Freiburg University Medical Center, Germany;* [e]*AGFA Healthcare, Ghent, Belgium*

Abstract. In this paper, we introduce a data integration methodology that promotes technical, syntactic and semantic interoperability for operational healthcare data sources. ETL processes provide access to different operational databases at the technical level. Furthermore, data instances have they syntax aligned according to biomedical terminologies using natural language processing. Finally, semantic web technologies are used to ensure common meaning and to provide ubiquitous access to the data. The system's performance and solvability assessments were carried out using clinical questions against seven healthcare institutions distributed across Europe. The architecture managed to provide interoperability within the limited heterogeneous grid of hospitals. Preliminary scalability result tests are provided.

Keywords. Data Integration, Interoperability, Semantic Integration, Ontology

1. Introduction

The last ten years have been marked by the most important increase of biomedical information in human history. Electronic health records cover a growing part of these data, ranging from clinical findings to genetic structures. However, secondary data usage to improve healthcare quality and patient safety are very limited.

Several integration systems have been proposed to handle issues related to lack of technical standards and semantics among different data sources [1-3]. These systems provide methods to cope with data location and accessibility but do not necessarily manage data content and their semantics. Recently, with the advent of semantic web technologies, new data integration approaches using ontologies were proposed [4,5].

This paper introduces a three-layer ontology-driven data integration framework [5] that provides interoperability to heterogeneous storage systems. The methodology does not restrict data sources to an enforced common schema and the integration is done on-demand. The system called *virtual Clinical Data Repository* (vCDR) is being deployed and evaluated in a network of seven European hospitals in the DebugIT (Detecting and Eliminating Bacteria Using Information Technology) project [6].

The vCDR is used by decision support systems for data mining and monitoring tasks, especially at population level. Nevertheless, its pseudo-anonymized data allows unique identifiers to be linked back to actual patient information by authorized actors.

[1] Douglas Teodoro, University Hospitals of Geneva - Division of Medical Information Sciences, Rue Gabrielle-Perret-Gentil 4, 1211 Geneva, Switzerland; E-mail: douglas.teodoro@hcuge.ch

2. Methods

The vCDR architecture provides homogeneous real time view on the data sources, featuring common access mode, standard syntax and unified computer-interpretable semantics. In the healthcare field, for cross-border integration, the data warehouse approach [1] is not a viable solution. Data providers are not allowed to store patient data outside of their intranet domain due to ethical reasons. Furthermore, view integration [2] cannot be applied because operational databases (ODB) have to be protected from on-the-fly accesses to preserve system stability.

To solve the aforementioned constraints, the vCDR is based on a hybrid ontology-driven integration approach [5], where multiple semantically flat data description ontologies (DDO) are mapped to a common semantically defined DebugIT Core Ontology (DCO) and its extending operational ontologies (OO) [7]. As shown in Fig. 1, the system focuses on three levels of conceptual interoperability [8]: technical (network protocol, database), syntactical (terminology) and semantic (knowledge formalization).

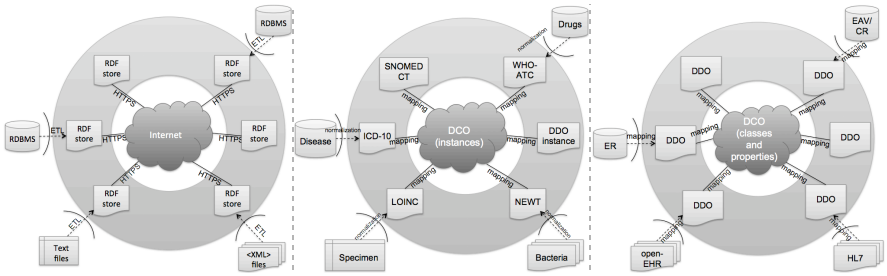

Figure 1: Three levels of interoperability in the integration platform - Technical (left) illustrating ODB standardization via SPARQL protocol and RDF storage; syntactic (center) illustrating the unification of site dependent values with terminologies and DCO instances; and semantic (right) illustrating how a description logics rooted formal ontology allows for DDO content unification and verification.

2.1. Technical interoperability

Clinical Information System (CIS) ODBs include different database management systems and access protocols. To provide a homogenous access layer, an intermediate storage is introduced between the CIS and the query point (Fig. 1 - left). The connection between the CIS and this local mirror - so called *local CDR* (lCDR) - is fulfilled by periodic Extract-Transform-Load (ETL) processes, which retrieve the content from the CIS, perform model transformations and load the data into the lCDR.

An lCDR comprises an RDF-like storage, usually backed by a relational database (RDB), featuring SPARQL communication protocol [7]. Numerous relational-data to RDF middleware approaches are proposed in the literature [9,10]. Despite not addressing the data integration problem, D2R [11] was chosen because it relies on the underlying RDB indexes to formulate the query plan, which gives better performance and scalability when compared to approaches that use native triple stores.

2.2. Syntactic interoperability

The content of DebugIT data sources are expressed in several languages and usually using free text. Thus, spelling mistakes and abbreviations such as *Staphyloccocus aureus* and *S. aureus* are commonly found. In order to bring syntactic alignment to the

lCDRs, their contents were transformed into a common syntax defined by biomedical terminologies (SNOMED CT, WHO-ATC, NEWT, etc.). These terminologies are mapped to DCO (terminology-to-DCO) using the SKOS ontology and Notation3 rules.

Specialized text mining algorithms were developed to perform term normalization [12,13] depending on the instance type. For example, for pathogen instances, the algorithm first tries to match the NEWT terminology against species, then against genus only. For antibiotics, it first tries to match the complete drug name against the WHO-ATC terminology, then the truncated 5-letters name. Finally, instances with small enumerated lists as value ranges were mapped manually.

2.3. Semantic interoperability

To bridge the gap between operational data and formal representations of concepts, the lCDR information model is formally defined using OWL language [7] to create a site-specific DDO. Moreover, shared representations of the domain concepts are derived to cover the clinical domain (DCO) and additional domains (OO) such as units, maths, hypothesis-generation, etc. Finally, links between the formal data source representations and the domain concepts are made through ontological mappings implemented via the SKOS ontology using the Notation3 format (DDO-to-DCO).

The SPARQL query language allows graphs to be built (*"construct"* clause) with DCO concepts using DDO terms in the *"where"* clause. Thus, a Global-as-View (GaV) approach (global ontology as view on the local ontology) can be applied in order to mediate data over the SPARQL endpoints of the lCDRs. For example, the query *"What is the resistance to <antibiotic> of <bacteria> during <period> at <location >?"* is translated as

```
CONSTRUCT
 {      ?antibiogram a dco:AntimicrobialSusceptibilityTest;
        biotop:hasAgent ?antibiotic; biotop:hasParticipant ?bacteria;
        biotop:hasOutcome ?outcome; dco:hasDate ?date.                  }
                                WHERE
{  DDO_SOURCE_1  }          {  DDO_SOURCE_2  }          {  DDO_SOURCE_N  }
```

with each DDO_SOURCE clause representing a lCDR query based on DDO terms.

It is during the query translation process provided by the *"construct"* algorithm that DDO concepts are annotated with DCO classes and properties. Binding variables are further converted using the terminology-to-DCO mappings provided in the syntactic alignment layer. Once this is done, the results are fully represented in terms of a formal ontology and their semantics are hence readily exploitable by computers.

3. Results

Seven healthcare institutions - GAMA (Sofia-BG), HUG (Geneva-CH), INSERM (Paris-FR), IZIP (Prague-CZ), LiU (Linköping-SE), TEILAM (Lamia-GR) and UKLFR (Freiburg-DE) - collaborated to evaluate the approach. They shared pseudo-anonymized historical episodes of care information, aggregated on unique identifiers of pathogens, thus avoiding patient-centric views.

In order to assess the system integration capability, i.e. which sites are able to answer clinical queries, and performance, i.e. how long it takes to retrieve a result set, in real life use-cases, the query "*What is the evolution of <bacteria> resistance to <antibiotic> during <period> at <location>?"* was exercised against the vCDR.

Fig. 2 shows the result of above query for *Pseudomonas aeruginosa* and *ciprofloxacin* in the *last 48 months up to Jun 2009* in the different hospitals. The system was able to obtain results from five out of seven institutions. The aggregated "DebugIT antibiogram" trend is shown in blue. Two of the lCDRs were not able to answer the query due to its constraints (*antibiotic*, *bacteria* and *period*).

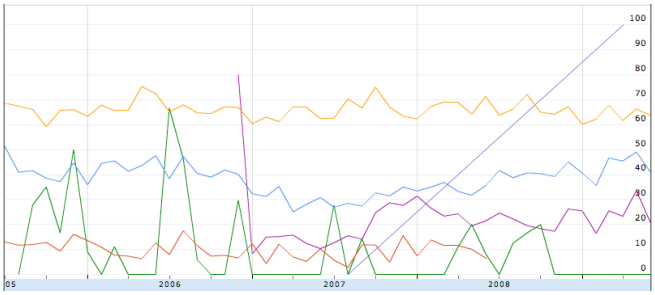

Figure 2: *P. aeruginosa* vs. *ciprofloxacin* resistance rate - Results shown here are not clinically relevant but rather useful to exercise the vCDR and were intentionally unlabelled to conform to hospital requirements.

To evaluate the performance of the SPARQL queries against the lCDRs, we executed the aforementioned query for *Klebsiella pneumonia* matching any *antibiotic* in order to increase the result set. Results presented in Table 1 show that network time is responsible for 41% to 49% of the retrieval time for the sets containing more than 1000 tuples. Indeed, due to their early stage of development, most SPARQL engines lack in aggregation functions such as *group by* and *count,* increasing the retrieval time.

Table 1: vCDR performance - The total time is the sum of the SPARQL engine time plus the network time. IZIP does not contain microbiology test results and TEILAM and GAMA have only a limited sample set.

Source	#Tuples Retrieved	Retrieval time (s)		#Tuples/sec
		SPARQL	Network	
HUG	74150	5.72	3.91	7704
INSERM	330360	20.38	14.22	9550
LIU	9905	1.70	1.23	3371
UKLFR	155315	6.34	6.19	12394

Finally, we compared the performance of the HUG SPARQL query presented in Table 1 with an equivalent SQL query using a direct access to HUG's RDB. The SQL query was executed in total 3.52s, which reduced the query time by 63%.

4. Concluding Remarks

The proposed vCDR architecture provides a three level integration framework. It is important to note that the approach deals with interoperability at each layer. Currently, data integration cannot be fully achieved with only the third layer of the proposed methodology, particularly for the case of operational databases.

The inexistence of global data model facilitates the seamless integration of new sources and ensures scalability. New data sources are only required to have a SPARQL endpoint formally described by a DDO and normalized instances. The domain ontology is not affected with the introduction of a new source. Instead, new terminology- and DDO-to-DCO mappings need to be created to represent each source added.

The syntactic alignment has shown to be a very complex process. The existence of linguistic and data type variances make it very difficult to find a common syntax; hence

the need for advanced natural language processing normalizers such as SNOCat [12]. The problem becomes even worse if intrinsic differences in defining “normal” values and thresholds are taken into account. For example, the measure for pathogen sensitivity to antibiotics is computed differently from country to country. The presence of a local expert is of utmost importance in these cases.

So far, the semantic integration is being extensively used without source model transparency. The final solution is a semantic mediator that allows users and query builders to select ontologically constrained idioms for query building. A proof of concept implementation is in an early stage. A previous version of a mediated vCDR was already described in [14]. In that version, besides the efficiency of the system in accomplishing the integration task, the constraint of a common unique schema has shown to be very restrictive to the project needs.

In this paper, an ontology-driven integration framework has been described. The architecture provides interoperability at technical, syntactic and semantic levels for heterogeneous clinical data sources. The system was assessed in a limited grid of seven EU healthcare centers. Despite an increase in the response time compared to traditional methods, the vCDR was able to retrieve results for a pre-defined set of queries in a satisfactory time for the project. The next step is the finalization of the semantic mediator contributing to increased end user compliance. Moreover, we plan to extend the syntactic aligner to a flexible framework to directly serve terminological servers and ontology look up services such as those maintained by epSOS, ECDC or the EBI.

Acknowledgements

This research has been supported by the EU-IST-FP7 DebugIT project # 712139.

References

[1] Shah SP, et al. Atlas - a data warehouse for integrative bioinformatics. BMC Bioinformatics. 2005;6: 34.
[2] Davidson SB, Overton GC, Tannen V, Wong L. BioKleisli: A Digital Library for Biomedical Researchers. International Journal on Digital Libraries. 1997;1:36-53.
[3] Stevens R, Baker P, Bechhofer S, Ng G, Jacoby A, Paton NW, Goble CA, Brass A. TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics. 2000;16(2):184-186.
[4] Shironoshita EP, Jean-Mary YR, Bradley RM, Kabuka MR. semCDI: a query formulation for semantic data integration in caBIG. J Am Med Inform Assoc. 2008;15:559-568.
[5] Cruz I, Xiao H. Ontology driven data integration in heterogeneous networks. Complex Systems in Knowledge-based Environments: Theory, Models and Applications. 2009:75-98.
[6] Lovis C, Colaert D, Stroetmann VN. DebugIT for patient safety - improving the treatment with antibiotics through multimedia data mining of heterogeneous clinical data. Stud Health Tech Inform. 2008;136:641
[7] Schober D, Boeker M, Bullenkamp J, Huszka C, Depraetere K, Teodoro D, et al. The DebugIT Core Ontology: semantic integration of antibiotics resistance patterns. Proceedings of MEDINFO 2010; Cape Town; 2010.
[8] Tolk A. What Comes After the Semantic Web - PADS Implications for the Dynamic Web. Proceedings of the 20th Workshop on Principles of Advanced and Distributed Simulation (PADS'06); 2006.
[9] Broekstra J, Kampman A, Van Harmelen F. Sesame: A generic architecture for storing and querying RDF and RDF schema. Proceedings of The Semantic Web - ISWC 2002; 2002. p. 54-68.
[10] Erling O, Mikhailov I. Virtuoso: RDF Support in a Native RDBMS. Semantic Web Information Management; 2010. p. 501-519.
[11] Bizer C, Cyganiak R. D2RQ Lessons Learned. W3C Workshop on RDF Access to Relational Databases; 2007.
[12] Ruch P, Gobeill J, Tbahriti I, et al. Automatic Assignment of SNOMED Categories: Preliminary and Qualitative Evaluations. First Semantic-Mining Conference on SNOMED CT – SMCS; 2006.
[13] Daumke P, Enders F, Simon K, et al. Semantic Annotation of Clinical Text - the Averbis Annotation Editor. Proceedings of the GMDS 2010; Mannheim, Germany.
[14] Teodoro D, Choquet R, Pasche E, et al. Biomedical Data Management: a Proposal Framework. Stud Health Technol Inform. 2009;150:175-9.