Semantic Interoperability of Clinical Data

Jacqueline Midlej do Espírito Santo and Claudia Bauzer Medeiros

Institute of Computing
University of Campinas - UNICAMP, BRA
{jacqueline.santo, cmbm}@ic.unicamp.br

Abstract. The interoperability of clinical information systems is particularly complicated due to the use of outdated technologies and the absence of consensus about standards. The literature applies standard-based approaches to achieve clinical data interoperability, but many systems do not adopt any standard, requiring a full redesigning process. Instead, we propose a generic computational approach that combines a hierarchical organization of mediator schemas to support the interoperability across distinct data sources. Second, our work takes advantage of knowledge bases to be linked to clinical data, and exploit these semantic linkages via queries. The paper shows case studies to validate our proposal.

Keywords: semantic interoperability, mediator, terminology, ontology

1 Introduction

The ability to exchange data among computational systems is called *data interoperability* [1]. Systems (and data) interoperability has been studied for decades, but there is still much to be done. In medical information systems, one of the most adopted solutions is the use of data exchange standards. However, the constant appearance of new devices and collection methods has exponentially increased the data heterogeneity problem. This paper addresses the problem of clinical data interoperability. Besides technical issues, clinical data interoperability is moreover hampered by ethical and security issues, by the absence of consensus about standards and terminology, and by the use of outdated and closed technology.

In the healthcare context, the two main approaches to clinical data interoperability are: (1) Electronic Health Records (EHR) standards to model clinical information and (2) terminologies to establish common understanding of vocabulary and concepts. EHR standads define protocols to digitally store and exchange patients' health data. The most frequently adopted standards are HL7[1], openEHR[2] and ISO/EN 13606[3]. However, many systems do not apply any standard - they have progressively adapted introduction of new technologies and would have to be completely redesigned to comply with standards.

[1] Health Level Seven. www.hl7.org

[2] OpenEHR Specification. www.openehr.org

[3] ISO/EN 13606 standard. www.iso.org or www.en13606.org

Research on the second approach (2) investigates the use of knowledge bases such as terminologies and domain ontologies. *Terminology* denotes all types of vocabularies, such as controlled vocabularies, thesaurus, code systems and so on. In clinical information systems, a terminology is used only to define a common understanding in the interoperability process, but does not explore any the semantics. The latter are achieved by use of ontologies to integrate several knowledge bases.

This paper presents a methodology for enabling the interoperability of arbitrary clinical information systems, exploring semantic aspects from knowledge bases instead of standards. As will be seen, this allows posing queries to data from arbitrary health centers, and also constructing new kinds of queries. Semantic interoperability allows relating facts that are not directly related, and knowledge discovery. Our proposal is backed up by a case study.

Section 2 presents the state-of-the-art on clinical data interoperability. Section 3 presents our approach. Section 4 shows two case studies using real data. Section 5 briefly presents initial prototype to query clinical data from distributed systems. Section 6 presents conclusions and future work.

2 Related Work

Table 1 summarizes some of the main research efforts on interoperability among clinical information systems. Column 2 indicates the interoperability strategy adopted. Columns 3-4 indicate which EHR standard is adopted, and expansibility options. Column 5 indicates whether medical terminologies are used.

Table 1. Related Work

	Interoperability Strategy	**Semantics**		
		EHR Standards		**Terminology**
		Standard	**Expansible**	
Jian [7]	single standard	TMT	Transformable into HL7	No
Li [8]	single standard	MML	To HL7. Other mappings can be made by user	Yes
Müller [10]	single standard	HL7	No	No
Hosseini [6]	single standard	HL7	No	Yes
Sartipi [12]	Mediator schema	HL7 + ACORD	Generic standard-based guideline	Yes
Azami [2]	mediator schema	any	Generic: non-standard-based	Yes
Costa [4]	Mediator ontology	ISO/EN 13606 + openEHR	To standards based on dual-model architecture	No
Berges [3]	Mediator ontology	any	Generic: any data model	Yes
Dogac [5]	P2P + mediator ontology	HL7 + ISO/EN 13606	Generic: any ontology model	No

Almost half of the reviewed papers uses a single EHR standard across many computer systems ([6–8,10]). Sometimes, the authors provide expansion to other standards ([7,8]). For instance, the Taiwan Electronic Medical Record Template (TMT) [7] can be transformed into HL7. Also, Li et al. [8] present rules to convert MML[4] versions and MML into HL7 to provide data exchange across countries.

The use of data standards may not be enough when systems are distributed and do not have the same data model. The mediator architecture is a classic

[4] Medical Markup Language. `www.medxml.net/E_mml30/mmlv3_E_index.htm`

strategy for interoperability of systems based on different data models. This approach defines a global mediator database schema and mappings between the mediator and local schemas. For example, Sartipi and Dehmoobad [12] propose a standard-based guideline to semantic interoperability in health subdomains, with HL7 to represent the information model in the clinical domain and the ACORD[5] standard for the insurance domain. The common part of the domain models (basically, their "intersection") produces the mediator model. On the other side, Azami et al. [2] do not define a single schema, but rather a set of hierarchical mediator schema to integrate health subdomains.

Berges et al. [3] and Costa et al. [4] also define a mediator model, but use the ontological representation of the data models. Costa et al. [4] define a common ontology for EHR standards based on a dual-model architecture (e.g., ISO/EN 13606 and openEHR). Berges et al. [3] use a generic approach independent from any standard. Local ontologies are semi-automatically generated from local repositories and are specializations of the common ontology.

Another approach is shown by Dogac et al. [5], who use Peer-to-Peer (P2P) architectures combined with a mediator architecture to create an ontological representation for each peer. It is a standard-based approach in which HL7 and ISO/EN 13606 are used to add meaning to the exchange process.

Besides the structural model, EHR standards define semantics for clinical information via conceptual model (e.g., archetypes defined in openEHR). The use of knowledge bases is another way to obtain semantics. In general, authors only use medical terminologies to define a common vocabulary. For example, Sartipi and Dehmoobad [12] propose a shared terminology system to add semantics to exchanged messages. A different application is to help the translation process, such as Li et al. [8]. However, unlike us, they do not use domain ontologies and do not explore the semantic relationships from these knowledge bases.

There are several integrated knowledge bases that can be explored in the health domain. For example, the UMLS[6] meta-thesaurus establishes links to 200 biomedical vocabularies. Also, the LODD [11] links data about drugs and the TMO [9] is a medical unified ontology. All these approaches model terminologies in an ontology language, allowing to link them to other domain ontologies.

3 Our architecture for Semantic Interoperability

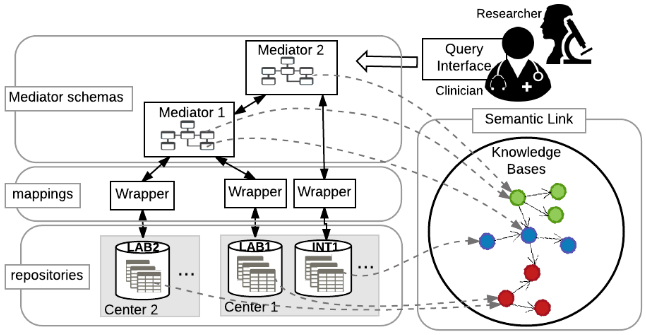
Figure 1 shows our architecture in two facets: a) on the left side, a mediator-based structure for clinical data interoperability and b) on the right side, semantic linkage with knowledge bases. These facets allow integrated access to clinical data, designing queries driven by clinical context and exploring semantics on query processing.

The left side is a hierarchical mediator architecture based on [2]. The bottom layer displays heterogeneous and distributed clinical data repositories from

[5] Association for Cooperative Operations Research and Development. `www.acord.org`

[6] Unified Medical Language System. `www.nlm.nih.gov/research/umls/`

Fig. 1. Architecture overview



several health centers. We do not assume any standard about the center's data models. Ours is a generic and non-standard based proposal.

The mediator layer presents a hierarchy of mediator schemas to allow integration of subdomains of health information systems. The mediators at the lower levels define commons schemas to integrate each subdomain from all component health centers. In the figure, schemas from `Center 1` and 2 are integrated in `Mediator 1` (e.g., in laboratories). Other subdomains, such as Chemotherapy, can be integrated the same way. The mediators at the upper levels integrate the subdomains (using other intermediate mediators when needed) creating a global view of all repositories at the top of the hierarchy. The mediator schemas are designed in collaboration with domain experts.

The mappings layer stores the relationships between a mediator schema and those of individual repositories. For each local model, a wrapper sets the matching with the mediator schema, identifying sameness regardless of the structural and syntactic differences. Semantic links to knowledge bases can help find these correspondences between schemas, identifying different attributes linked to equivalent and/or even related terms. However, the repositories can have attributes that are not mapped. Usually, a mediator has a more global view that cannot cover all details. Also, more than one global mediator can be defined for different query purposes, e.g. clinical or research. So, besides the most common queries that are centered on a single patient, we foresee to query for sets of patients with similar symptoms Section 4.1 exemplifies the latter by setting the context by specifying results of tests, diagnostics, medications, etc.

This first facet allows users to recover data from different healthcare centers in the following way: 1) Users state queries based on the mediator schema; 2) Queries are addressed to intermediate mediator schemas and are reformulated according to mapping rules encapsulated in wrappers, creating subqueries consistent with the local schemas; 3) Subqueries are forwarded to the local models, which return the results back to the wrapper; 4) Wrappers unify the results according to the mediator schemas; 5) At the end, the results from all repositories are returned to the user interface.

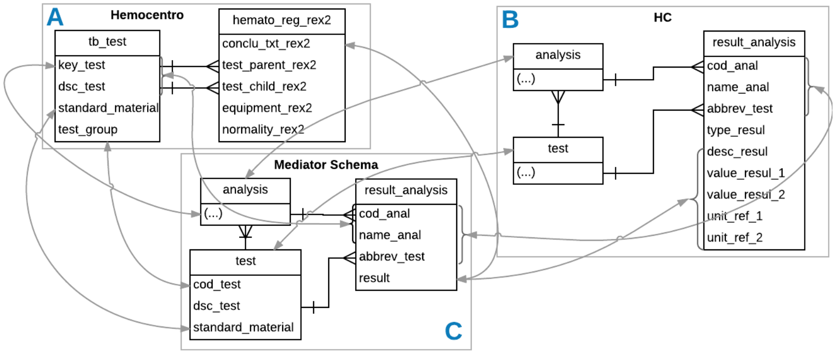
The main difference between our proposal and standard mediator architectures is the possibility of queries using semantic links across clinical data sources and integrated knowledge bases (such as LODD [11] and TMO [9]). This second facet, on the right side of figure 1, considers links to knowledge bases from instances and attribute labels of the schemas. Various types of relationships can be exploited, e.g.: equivalence, composition, source, or causality. They can also be extracted from the existing relationships in the repositories. These semantic relationships can be exploited in query expansions - see section 4.2.

4 Case Study

4.1 Interoperability of Laboratory Test Results

Let us now exemplify the mediator approach on real data from *Hospital das Clínicas*(HC) and *Hemocentro* databases at University of Campinas. HC is one of the biggest hospitals in Brazil, with 44 medical specialties and 5 thousand tests and exams performed per day. Its clinical and administrative information is distributed over 19 systems and many different DBMS. Hemocentro is a hemopathy specialized center, having a single information system.

Fig. 2. Schemas of Test Results



Parts A and B of Figure 2 show an excerpt of the schemas that define how results of laboratory tests are stored in each center. Part C presents a possible mediator schema, in which arrows show correspondences among the models. This example shows how lab tests are treated differently in Hemocentro and HC. E.g., `result` in the mediator schema is mapped to five attributes in HC schema. Also, attribute `test_group` in Hemocentro plays the same semantic role as the relationships between test and analysis in HC.

Consider now **Query1** for *Platelet Count* (an analysis of *Blood Cell Count*) submitted to the mediator. Following the arrows, wrappers formulate **Query2** and **Query3**, addressed to Hemocentro and HC schemas, respectively. After these queries are executed, the results is unified and send back to the user.

```
Query1
SELECT result
FROM result_analysis
WHERE name_anal='Platelet
Count';
```

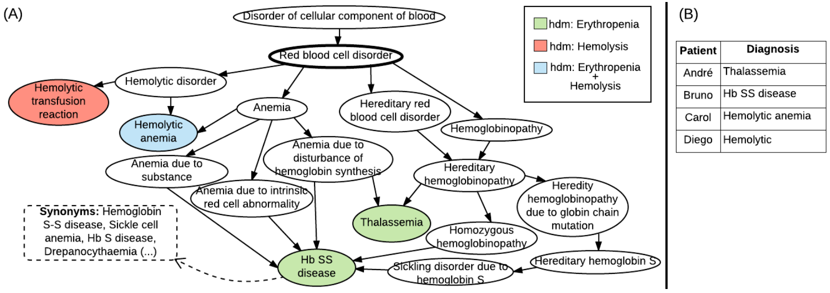
```
Query2
SELECT A.conclu_txt_rex2
FROM hemato_reg_rex2 as A,
tb_test as B
WHERE A.test_child_rex2=B.key_test
AND B.dsc test='Platelet Count';
```

```
Query3
SELECT desc_resul, value_resul_1,
unit_ref_1, value_resul_2,
unit_ref_2
FROM result_analysis
WHERE name_anal='Platelet Count';
```

4.2 Semantic Enrichment in querying Blood Cell Disorders

This section shows queries exploring knowledge provided by terminologies and ontologies in three ways: 1) **Synonyms**: to search a concept expanding the search to all synonymous terms; 2) **Generic concepts**: to search a broader concept, also recovering results from the specific terms; 3) **Relationships**: to search a concept that has a specific relationship with another concept. Part A of figure 3 illustrates a cut of the terminology SNOMED CT. It is the bigger clinical terminology - the US edition[7] in 2015 has more than 300 thousand unique concepts and 900 thousand relationships between concepts. Part B shows hypothetical records about the diagnosis of four patients linked to SNOMED CT.

Fig. 3. Cut of SNOMED CT


(A)
Disorder of cellular component of blood
Red blood cell disorder
Hemolytic disorder
Hemolytic transfusion reaction
Hemolytic anemia
Anemia
Hereditary red blood cell disorder
Hemoglobinopathy
Anemia due to disturbance of hemoglobin synthesis
Anemia due to substance
Anemia due to intrinsic red cell abnormality
Hereditary hemoglobinopathy
Heredity hemoglobinopathy due to globin chain mutation
Thalassemia
Homozygous hemoglobinopathy
Synonyms: Hemoglobin S-S disease, Sickle cell anemia, Hb S disease, Drepanocythaemia (...)
Hb SS disease
Sickling disorder due to hemoglobin S
Hereditary hemoglobin S
🟩 hdm: Erythropenia
🟥 hdm: Hemolysis
🟦 hdm: Erythropenia + Hemolysis
(B)

Patient	Diagnosis
André	Thalassemia
Bruno	Hb SS disease
Carol	Hemolytic anemia
Diego	Hemolytic

To exemplify way (1), SNOMED CT lists synonyms for each concept. The figure shows the synonyms for `Hb SS disease`. Any term listed can be used to retrieve patients who have this disease. For example, when querying patients who have `Sickle cell anemia` the result would be Bruno.

To exemplify way (2), the arrows show `is_a` relationships. A query using a more generic term can return results from all its more specific terms by inference. For example, the query for patients who have some `red blood cell disorder` returns all patients, while the query for patients who have some `hereditary red blood disorder` returns André and Bruno.

To exemplify way (3), SNOMED CT defines other relationships via attributes. The colors show values given to attribute `has_definitional_manifestation` (`hdm`): `Erythropenia` and `Hemolysis`. According to SNOMED CT, `hdm` "links disorders to the manifestations (observations) that define them". Erythropenia means the number reduction of red cells and Hemolysis means the premature

[7] Systematized Nomenclature of Medicine Clinical Terms. `www.nlm.nih.gov/research/umls/sourcereleasedocs/current/SNOMEDCT_US/index.html`

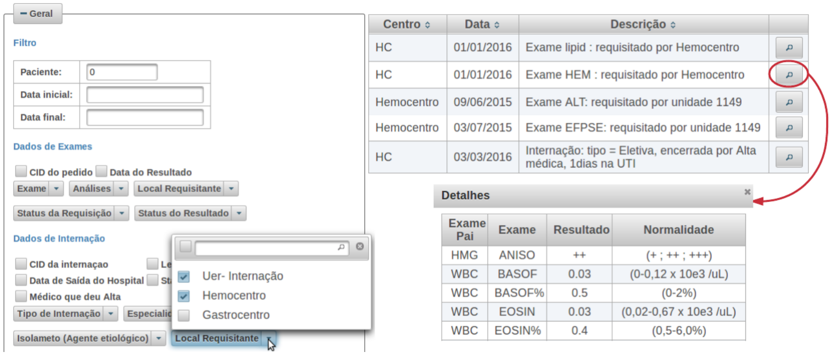
destruction of red cells. A search for patients who has `Erythropenia` returns Carol and Diego. A query for patients with `Hemolysis` returns André, Bruno and Carol.

The combination of semantics and the mediator architecture allows expanding queries at each component center, thus providing a wide range of possibilities for research and treatment.

5 Prototype

Figure 4 shows a screen copy of first version of the prototype (in Portuguese) to perform clinical data interoperability among three databases: laboratory tests at Hemocentro and HC, and hospitalization at HC. This version is not yet semantically linked to knowledge bases. The left side is an interface to design queries, filtering by attributes. It allows users to specify the health context of interest. The right side shows results of a patient's medical history.

Fig. 4. Prototype interface



Paciente: 0

Data inicial:

Data final:

☑ Uer- Internação

☑ Hemocentro

□ Gastrocentro

Centro	Data	Descrição
HC	01/01/2016	Exame lipid : requisitado por Hemocentro
HC	01/01/2016	Exame HEM : requisitado por Hemocentro
Hemocentro	09/06/2015	Exame ALT: requisitado por unidade 1149
Hemocentro	03/07/2015	Exame EFPSE: requisitado por unidade 1149
HC	03/03/2016	Internação: tipo = Eletiva, encerrada por Alta médica, 1dias na UTI

Exame Pai	Exame	Resultado	Normalidade
HMG	ANISO	++	(+ ; ++ ; +++)
WBC	BASOF	0.03	(0-0,12 x 10e3 /uL)
WBC	BASOF%	0.5	(0-2%)
WBC	EOSIN	0.03	(0,02-0,67 x 10e3 /uL)
WBC	EOSIN%	0.4	(0,5-6,0%)

6 Conclusions and Ongoing Work

The paper proposes a framework to achieve semantic interoperability of clinical data in health information systems. We introduce a non-standard approach composed of two steps: 1) clinical data interoperability and 2) semantic enrichment. The first step uses a hierarchy of mediators to integrate distributed systems from different healthcare centers. The first case study exemplifies this step using real schemas from Unicamp healthcare complex. However, this example still faces problems about the use of different vocabularies, solved by semantic linkage.

This linkage brings new ways to explore clinical data, helping knowledge discovery. The second case study unifies the vocabulary and presents queries using SNOMED CT. Although most health centers use International Classification Disease (ICD) codes, it does not set semantic relationships between concepts.

Future work includes the expansion of the clinical subdomains, covering most computer systems at HC and Hemocentro. Our main effort is to establish links to integrated life science terminologies and ontologies and to explore this knowledge. Another promising direction involves ontology/terminology evolution, which will require dynamically refreshing system links and information.

Acknowledgements Work partially financed by 1504650 (CAPES), 142243/2017-5 (CNPq), FAPESP/ Cepid in Computational Engineering and Sciences (2013/ 08293-7), INCT in Web Science. Our thanks to Hemocentro and HC collaborators, mainly Edson Kitaka, Profs. Fernando Costa and Erich de Paula.

References

1. IEEE standard computer dictionary: A compilation of IEEE standard computer glossaries. IEEE Std 610 pp. 1–217 (Jan 1991)
2. Azami, I., Malki, M., Tahon, C.: Integrating hospital information systems in healthcare institutions: A mediation architecture. J. Med. Syst. 36(5), 3123–3134 (Oct 2012)
3. Berges, I., Bermudez, J., Illarramendi, A.: Toward semantic interoperability of electronic health records. IEEE Transactions on Information Technology in Biomedicine 16(3), 424–431 (May 2012)
4. Costa, C., Menárguez-Tortosa, M., Fernández-Breis, J.: Clinical data interoperability based on archetype transformation. J. of Biomedical Info. 44(5), 869–880 (Oct 2011)
5. Dogac, A., Laleci, G., Kirbas, S., Kabak, Y., Sinir, S., Yildiz, A., Gurcan, Y.: Artemis: Deploying semantically enriched web services in the healthcare domain. Information Systems 31(4–5), 321–339 (2006)
6. Hosseini, M., Ahmadi, M., Dixon, B.: A service oriented architecture approach to achieve interoperability between immunization information systems in Iran. AMIA Annual Symposium Proc. pp. 1797–1805 (Nov 2014)
7. Jian, W., Hsu, C., Hao, T., Wen, H., Hsu, M., Lee, Y., Li, Y., Chang, P.: Building a portable data and information interoperability infrastructure - framework for a standard Taiwan electronic medical record template. Computer Methods and Programs in Biomedicine 88(2), 102–111 (2007)
8. Li, J., Zhou, T., Chu, J., Araki, K., Yoshihara, H.: Design and development of an international clinical data exchange system: the international layer function of the Dolphin project. J. of the American Med. Info. Association 18(5), 683–689 (2011)
9. Luciano, J., Andersson, B., Batchelor, C., et al.: The translational medicine ontology and knowledge base: driving personalized medicine by bridging the gap between bench and bedside. J. of Biomedical Semantics 2(suppl 2), 1–21 (2011)
10. Müller, M.L., Ückert, F., Bürkle, T., Prokosch, H.U.: Cross-institutional data exchange using the clinical document architecture (CDA). Int. Journal of Medical Informatics 74(2-4), 245 – 256 (2005)
11. Samwald, M., Jentzsch, A., Bouton, C., Kallesøe, C., Willighagen, E., Hajagos, J., Marshall, M., Prud'hommeaux, E., Hassanzadeh, O., Pichler, E., Stephens, S.: Linked Open Drug Data for pharmaceutical research and development. J. of Cheminformatics 3(1), 19 (2011)
12. Sartipi, K., Dehmoobad, A.: Cross-domain information and service interoperability. In: iiWAS. pp. 25–32. ACM, New York, NY, USA (2008)