Exploring Semantics in Clinical Data Interoperability

Jacqueline Midlej do Espírito Santo[1], Erich Vinicius de Paula[2], and Claudia Bauzer Medeiros[1]

[1] Institute of Computing - University of Campinas - UNICAMP, BRA
{jacqueline.santo, cmbm}@ic.unicamp.br
[2] Faculty of Medical Sciences - University of Campinas - UNICAMP, BRA
{erich}@unicamp.br

Abstract. The increasing amount of digital clinical information has prompted research in interoperating across numerous clinical data sources. Most solutions to this problem follow one of two main directions: (a) adoption of Electronic Health Records (EHR) standards, or (b) structuring medical knowledge via Knowledge Organizations Systems (KOS). Related research sometimes addresses the combination of the two directions but does not explore the knowledge of KOS, which are just used to define and disambiguate concepts. This paper discusses the solution for clinical data interoperability that we designed and implemented. It is a two step process - (a) we provide initial integration via mediators to provide mappings across heterogeneous sources, and (b) KOS to extend navigation possibilities across data sources. This paper is centered in the second part, discussing the challenges of semantic query expansion for clinical data analysis. We illustrate our solution through a real case study from one of Brazil's largest hospital complexes.

Keywords: interoperability · Medical Knowledge Organizations Systems · semantic query · query expansion

1 Introduction

This paper is concerned with interoperability challenges in eHealth[3], in particular those associated with clinical data management. Two directions have been taken to facilitate clinical data interoperability: (a) adoption of Electronic Health Records (EHR) standards or (b) structuring medical knowledge via Knowledge Organizations Systems (KOS). EHR standards are specifications about how medical data should be structured and stored to facilitate interoperability among different health systems. Their adoption often requires extensive recoding. We, instead, adopt the classical mediator strategy to deal with different clinical information systems regardless of EHR standards. The other prevalent solution to interoperability is based on KOS. The term KOS is intended to encompass

[3] Here defined as a multidisciplinary research field that requires collaboration of computer scientists with researchers in HealthSciences.

all types of schemes for organizing information and promoting knowledge management, such as dictionaries, taxonomies, thesauri, and ontologies [5]; the two latter are the most common KOS used to semantically organize clinical data. There are hundreds of medical KOS; some of them are *de facto* standards (such as the International Classification of Diseases - ICD) but most have no consensual use. Research on clinical data interoperability that involves KOS can be classified in two main directions: the first uses them to disambiguate terms for interoperability, but does not take advantage of the power of ontologies to expand queries; the second constructs (usually small) case-specific ontologies to expand queries, thereby helping find new facts in a specific clinical (sub)domain.

We, instead, use existing generic KOS to expand queries, thereby generalizing the second approach to arbitrary clinical information systems. To the best of our knowledge, ours is the first proposal to combine mediators to KOS to expand queries, thereby helping users query and explore data in clinical information systems. The term *users*, in this text, refers to health professionals that work and perform research in primary or secondary health care (e.g., doctors or nurses). Indeed, our approach addresses the interoperability of arbitrary clinical information systems, regardless of ERH standards, exploring semantic aspects by navigating integrated medical KOS.

We showcase our approach through a real case study that illustrates the challenges of extracting, from a large set of heterogeneous clinical data sources, *ad hoc* patient groups for subsequent analysis. This scenario is typical of the demands of clinical research, but is also found in situations where, e.g., hospital administrators need to analyze costs associated to a given set of pathologies. Our architecture was published previously in [3], where we restricted ourselves to the mediator aspects, but did not discuss semantic query expansion. Thus, our main contributions are: a) a new approach to support semantic queries over arbitrary clinical information systems; and b) a discussion of challenges in a real scenario exploring knowledge extracted from KOS to help users in query formulation.

Part of the complexity of our work lies is the complexity of our clinical scenario and associated data, which is typical of many clinical systems in which legacy data have to live with new systems and data collecting devices. Our work is being validated in a real, big data, health environment - one of Brazil's largest medical complexes, located at the University of Campinas (UNICAMP), Brazil. Clinical care in UNICAMP dates back to the 60's. The hospital systems rely on 19 different databases, each of which with tens of tables, with hundreds of attributes, and under distinct DBMS. Besides the hospital itself, the medical complex has 4 large specialized health centers each with its own independent systems and data, and do not interoperate. Our testbed comes from two distinct centers inside this complex: Hospital das Clínicas (HC) and Hemocentro. Hemocentro is a center of hematology and hemotherapy that treats 1500 patients/month and manages 70 thousand blood donations/year. The hospital has 44 medical specialties, performs about 5 thousand laboratory test/day and 15 thousand hospitalizations/year. Our tests are being conducted on a 5 year extract of these data, for approximately 40 thousand patients.

2 Related Work

Our approach to interoperability in clinical systems combines the use of mediators with semantics provided by KOS to expand queries. We discussed our mediator approach in [3]. Thus, this section focus on the semantic query formulation process, including query expansion. Given the specificity of the clinical domain, most of our references relate to the use of ontologies in this field.

Semantic queries can be defined as queries that leverage the semantic information stored in ontologies to filter and retrieve data from relational tables [8]. In the health context, data from clinical centers are mostly stored in relational databases while other medical information can be found in ontology models, spreadsheets or textual documents. Semantic annotations are used to establish the linkage between ontologies and relational data. Similarity functions can be used to find their correspondences. We adopt the definition of [9] of semantic annotations, as follows: "Semantic annotations combine concepts of metadata and ontologies: metadata fields are filled with ontology terms, which are used to describe these fields. A semantic annotation unit is a triple <s,m,o>, where s is the subject being described, m is the label of a metadata field and o is a term from a domain ontology." We use Bioportal [14] to annotate clinical data. Bioportal is a repository with 768 integrated biomedical ontologies and offers a REST API and SPARQL endpoint to access this repository programmatically.

One can organize the query formulation process, from a high point of view, in a sequence of interconnected phases: initial query formulation, query reformulation, and query processing. The initial query formulation can appear under different guises, which can be roughly classified into (a) direct formulation (the user writes the query in some sort of language), and (b) interactive formulation (a query system, e.g. in [8], guides users into expressing the query via record patterns). Query reformulation basically consists in, given a query in some language, rewriting it - in the same, or another language - to achieve some kind of goal (e.g., extending results [2, 13, 15–17] or semantic interpretation [2, 16]). Finally, query processing involves the execution of the reformulated query. These phases can be repeatedly executed until the user is satisfied with the result.

Table 1 (a) summarizes related work, identifying in which phase the work has their main contribution, the most common goals to work in some query formulation phase, the role of ontologies in the process and the domain of the research and case study. Table 1 (b) shows some related works.

Related work centered on the initial query formulation phase usually provides solutions to facilitate query construction. Most papers propose a query interface to guide the user in the query formulation process, either by navigating through concepts in the ontology (e.g. [1,12]), helping the user to specify search conditions in a graphical way (e.g. [10]), or proposing a query language (e.g. [7]).

Though, of course, an initial formulation is required, we are more interested in the reformulation phase. According to Vilar [13], reformulation can be found under different names, each of which denotes some kind of algorithmic rewriting technique - semantic rewriting, syntactic rewriting, expansion. These classifications vary from author to author. For instance, semantic rewriting is often

Table 1: Classifications of related work in query formulation process

(a) Query formulation process

Criterion	Approach
Phase	P1. Initial query formulation P2. Query reformulation P3. Query processing
Goal	G1. Facilitate the use (in P1) G2. Data integration (in P2) G3. Obtain extended results (in P2) G4. Semantics interpretation (in P2) G5. Optimization (in P3)
Ontology	O1. Application data model O2. Additional domain information
Domain	Generic, Health or Other

(b) The related work

Work	Phase	Goal	Ontology	Domain
Boonprapasri [1]	P1	G1	O1	Other: GIS
Lelong [7]	P1	G1	O2	Health: EHR
Tiede [12]	P1	G4	O2	Other: Geography
Munir [10]	P1 P2	G1	O1 O2	Health
Zheng [17]	P2	G3	O2	Heath: Biomedicine
Yunzhi [15]	P2	G3	O2	Health: Hepatitis
Zhao [16]	P2	G3 G4	O2	Health: Image note
Vilar [13]	P2	G3	O1 O2	Other: Biology
Calvanese [2]	P2	G2 G3 G4	O1	Generic

intended to obtain distinct results (either more generic or more specific). Hence, some authors do not consider this a reformulation, given that the results may not be identical to the original formulation.

Research that addresses query reformulation adopts one of the following strategies: 1) define operations over ontologies, applying the results in relational query language declarations or 2) transform the relational schema into an ontology and then address queries only via ontology processing. In the first strategy, the ontology always plays the role of bringing additional domain information into the application. In the second strategy, the ontology always plays the role of the application data model. This means that an ontology model is used to organize the data sets of the application. Both strategies can be applied together using multiple ontologies. For example, we can have one application ontology that corresponds to the specific data itself (e.g., modeling how results of laboratory tests are stored); and we can have multiple ontologies to bring additional knowledge about the tests and possible diagnoses (e.g., the range of reference values of a test, and diseases it can detect).

For example, Zheng, Wang and Lu [17] use the first strategy to define some operations used as an extension for relational query languages (such as getHyponym, getHypernym, getSynonym, and getSibling). The operations expand the query by adding new terms related to the original one, thus enabling to recover more information than the original query.

Vilar [13] and Calvanese [2] adopt the second strategy – that maps all relational schemas into ontologies. Vilar [13] performs two expansion options: the system finds existing domain ontologies that are potentially good for query expansion, or the users choose the expansions they want, using a predefined set of operations to navigate through the ontologies. While Vilar [13] adopts a relational query language extended with a set of operations (similarly to [17]), Calvanese et al [2] use an ontological query language and address the issue of mapping queries in an ontology model to relational data sources.

Query expansion is also widely used to recover documents lacking relational structure. For instance, Zhao et al [16] combine query expansion to Natural Language Processing (NLP) to retrieve reports concerning medical image studies. Yunzhi et al [15] create a Hepatitis ontology to expand the terms annotated in health publications. Rather than creating an ontology Sonntag and Moller [11] use an existing domain ontology for query expansion.

As will be seen, our approach follows the first strategy - we propose a sequence of steps instead of operations to expand queries, and add new terms to query declarations based on ontology navigation. Unlike the second strategy, we do not transform a relational schema into an ontology.

3 Combining mediators to semantic processing for clinical data interoperability

Figure 1 depicts our architecture for interoperability of clinical systems, catering to both precision medicine and medical research needs. It shows, on the left side, a classical mediator approach to integrating data from several health centers (details in our previous work [3]). In most cases, the information needed in clinical care can be obtained using only the left part of the architecture, since clinical care is strongly dependent on a patient's medical history, which can be recovered following our mediator approach. However, medical research requires more complex analyses, whose specification depends on the researcher's needs and vocabulary, which seldom matches schema definitions, or terms entered in clinical databases.

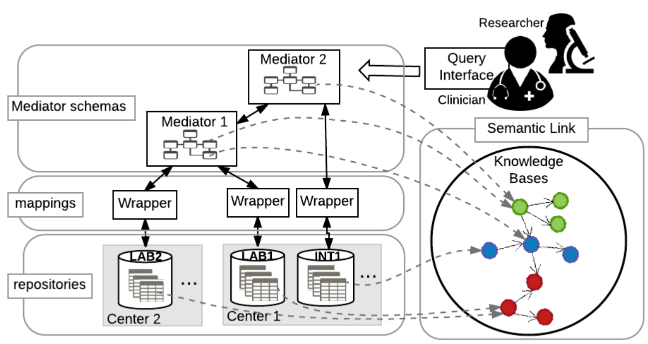

Fig. 1: Integration architecture - figure extracted from our previous work [3]

4 Adding Semantic Linkage

To address this issue, we use knowledge extracted from medical KOS, which we combine with the mediator integration approach. This section details how we explore semantics from the KOS and how they can help complex query processing and writing. This corresponds to the right part of figure 1.

Our semantic processing follows two stages: A) semantic annotation and B) query (re)formulation. Stage A concerns the creation of semantic annotations via inserting links between data and ontologies. Creation of these annotations requires finding the appropriate semantics for a term, and inserting the appropriate links. Semantics can be found by some similarity function that relates a term to ontology concepts. We use the annotator service in Bioportal REST API to find the appropriate semantics. It returns a set of ontology links for an input term (or a set of terms).

In many situations - such as our case study - creating one annotation per database term wastes storage space. In our sample data, HC performed more than 3 million lab tests, thus a given test name may appear in the HC database a few thousand times. Thus, we decided to store the semantic annotations themselves in a separate data table – our *Semantic Link Table – SLT* – to allow associating the same semantic annotation with many data records. Each entry in *SLT* is of form <database-term, id, url, ontology-name>, where "database-term" is, for instance, "hemogram". Afterwards, when needed, the semantic linkage is obtained via natural join of the data tables with SLT.

Once SLT is created, query reformulation (stage B) combines queries over medical ontologies with queries over our relational clinical data, as follows:

1. Query Bioportal to retrieve all ontological terms associated with a query term Q, given specific criteria (e.g., all descendants), obtaining a set $\{OT\}$. Each OT in $\{OT\}$ is a complex object containing all properties and relationships associated with Q.
2. For each OT, check if its id is in SLT.
3. If yes, add the corresponding SQL WHERE clause.
4. Once all clauses are written, pose the expanded query in the mediator-defined tables.

5 Case Study

We showcase our approach via a case study based on real, anonymized, data from Hemocentro and HC. The interoperability approach unifies datasets with five years (2012 to 2016) of information about laboratory tests, hospitalizations, and drug prescriptions for approximately 40 thousand patients, 13 million lab tests and 8 million drug administrations. We omit the mediator details (the left part of our architecture) and concentrate on how to take advantage of ontologies to process specific queries.

Every hospital in Brazil stores ICD codes associated with procedures, for billing information. However, the ICD hierarchy is exclusionary, not allowing the same disease to be part of several categories. Moreover, experts may consider disease categories not included in ICD classification. In any such case, the solution is to manually complement the ICD codes to create the context of interest. For example, in [6] the US Institute for Health Metrics and Evaluation. shows the ICDs codes selected to analyze the burden of diseases and causes of death in different categories. Some of them are not included in the ICD classification and, even considering the categories that are in ICD, they may include specific ICD codes outside the category.

Consider the following query *Retrieve all patients diagnosed with some Sexually Transmitted Disease (STD)*. There is no single ICD code for STD, and many such diseases are recognized via combinations of symptoms. Therefore, to process such a query, experts need to manually provide a combination of factors, a tiresome and error-prone task.

We now show how this can be performed via navigation through KOS. The following steps exemplify how to navigate in the Medical Subject Headings (MESH) via the Bioportal service to find the group of patients with STD (this correspond to steps 1 to 4 of section 4).

1. Given a term ("STD"), we create a chain of queries to Bioportal to: (a) recover the matching concept in MESH; (b) recover its descendants and (c) recover, for each descendant, it is mapping to the ICD classification.
2. Check if each ICD code is in SLT
3. Construct the SQL clause
4. Pose the query to the mediator-defined tables.

Focusing on ontology navigation, Part A of figure 2 shows an extract of the implementation, abstracting some details. Part a, b and c correspond to requests to Bioportal. In b, we obtain all MESH concepts included in the STD category (descendant concepts). However step c is needed because MESH terms are not directly linked to clinical databases, whereas ICD codes are in SLT. In 3, we show the expanded query in WHERE clause. Hiding details about the mediation process, part B shows a small excerpt of the result set: patients diagnosed with some STD.

A

```
    REST_URL="http://data.bioontology.org";
a.  meshConcept = jsonToNode(get(REST_URL + "/search?q=" + cat_name+
                        "&require_exact_match=true&ontology=MESH"));
b.  meshUrl=meshConcept.get(id);
    DescList = jsonToNode(get(REST_URL+ "ontologies/MESH/classes/" +
                  meshUrl+ "/descendants"));
c.  for(int i=0; i < descList.size(); i++)
      resultSet=jsonToNode(get(REST_URL+"ontologies/MESH/classes/"
                +descList(i)+"/mappings"));
      for (result : resultSet)
          if (result.contains("ICD10"))
              icdList.add(result.get(id));
3.  sql_std="SELECT patient, diagnostic FROM hospitalization WHERE" +
              "(diagnostic>=A50 AND diagnosti<=A64.9)";
    sql_expanded=sql_std+"OR diagnostic  IN  ("+ icdList +");" ;
```

B

Category: sexually transmitted diseases
Expanded ICD codes from MESH:
A50-A64.9, A71.9 A74.0, A74.9, B20-24.9, F02.4, I98.0, J16.0, Z21

Patient	Diagnostic
10332578	A539 D70
12419514	M869 A539
12381731	A86 B220
9932056	B24 B220
11436151	A55 A09 B201
7531282	Z21 L010
10926692	A58
12015978	N048 A530
9343015	S062 Z21

Fig. 2: Extracting MESH knowledge to help identifying patients with STD

In ICD classification, STD covers the code range A50-A64.9. Using MESH, we get more comprehensive results: (A50-A64.9), HIV range (B20-24.9) and 7 additional ICD codes, as shown in Part B of figure 2. To expand even more the results, we can follow the same process using additional KOS, such as SNOMED

CT and Disease Ontology. The standard query (without MESH expansion) for patients with STD only retrieved 95 patients from HC databases, while the MESH-expanded query retrieved 846 patients. Most of the latter were associated with some ICD code in the HIV range.

6 Discussion

Our full approach combines the use of mediators with semantic expansion. Mediators allow recovering individual patient data from heterogeneous data sources in a unified way, addressing precision medicine needs. Semantic expansion relies on the construction of our Semantic Link Table, SLT, and subsequent ontology navigation to help query formulation. Although KOS are widely adopted in the literature as part of interoperability strategies, they often play a minor role in clinical query reformulation. We highlight two benefits of exploring ontologies: 1) helping users formulate queries: since some information can be extracted from ontologies, they do not need to exhaustively describe the whole context; 2) formulate more complex queries and expand the results. These queries characterize clinical research needs, and are requested less often than those for clinical care needs.

In the case study we simplify the user's task by inferring diseases that belong to the STD category, instead of letting the user specify all such diseases. Although UNICAMP health centers link data using ICD codes, this is insufficient to, e.g., categorize diseases. ICD is a mutually exclusive and exhaustive statistically-based classification that creates arbitrary associations; when a concept should belong to two classes, and one is chosen, the underlying assumption is that it does not belong to another [4]. Also, search for patients within a given disease category is a complex query because it retrieves patients and diagnoses that are not directly annotated with the indicated query term. In the case study, the query term is "STD" and the result set includes patients with diagnoses that have been classified by their doctors as having, for instance, A55-Chlamydial lymphogranuloma or B24-Unspecified HIV disease, all of which were included in the STD category thanks to our approach (but which cannot be identified using an ICD-based approach alone).

Finding appropriate semantic concepts given a term is a challenging task. At HC this is especially aggravated because most of the data are entered using the hospital's internal codes – e.g., replacing test names by abbreviations. Automatic search for semantic links may return wrong associations. A semi-automatic approach increases link accuracy, for example letting the database designer validate the semantic links or choosing the main target ontologies. Another challenge in this approach is language translation. Brazilian health centers store data in Portuguese, while most medical KOS have no accredited Portuguese version. Therefore, an expert needs to validate the semantic links of our solution, otherwise we run the risk of creating wrong associations.

Another challenge is ontology navigation. Besides synonyms, narrower and broader terms, we can further explore the links between multiple ontologies and the other kinds of relations of a concept. For example, finding means to characterize a given disease via ontology navigating, such as to know its common symptoms, test results, and drug prescriptions. By doing this, we can infer additional diagnoses even if the name or ICD code is not written on the health centers database.

Last but not least, a combination of NLP and ontology linkage is yet another possibility to help query expansion, using the annotations entered by doctors on their patients. Unfortunately, our preliminary studies show that this created a large amount of false positives. We identified a non negligible number of records in which doctors' annotations identified both the illness, and explained discarded hypotheses. Thus, we would need to invest into more sophisticated NLP techniques to be able to retrieve meaningful records.

7 Conclusions and ongoing work

This paper discussed the challenges of exploring knowledge from ontologies for clinical data interoperability. In spite of extensive research in this field, solutions still tend to concentrate in a given trend – namely, use of standards and mediators, or adoption of semantics to enhance data understandability, without exploiting the full possibilities of ontological processing. To the best of our knowledge, ours is the first proposal that combines both trends in a generic, extensible architecture in which ontologies are used in query expansion. We exemplify how it can be used via real life, big data, test case from one of Brazil's largest medical compounds. The discussion of this case shows some of the many challenges faced in handling clinical data – from its intrinsic heterogeneity, even within a single dataset, to the dependence on non-consensual vocabularies and ontologies, and need for NLP. While some of these challenges are specific to our test environment (e.g., the particular characteristics of the systems and data we had to deal with), others are generic. Our case study helps to exemplify generic challenges, such as having to cope with legacy data and systems and different query requirement patterns, such as finding clusters to support research and decision making.

Our ongoing work involves both research and development activities. On the latter side, we are continuing our development efforts to include additional ontologies and vocabularies, and to check more complex situations. In this, we are being helped by medical experts to express requirements and validate (and question) results. On the research side, more needs to be done towards NLP. In our test context, some of the medical systems within our complex are being remodelled to support automated electronic health record handling, and support patient care "workflows". This, in turn, requires considering the evolution of data and versioning of database schemas, while at the same time supporting the legacy systems.

Acknowledgements Work partially supported by FAPESP/CEPID CCES under grant 2013/08293-7, and individual grants from CNPq.

References

1. Boonprapasri, T., Sriharee, G.: An applied ontology: A semantic query builder for health gis system. In: ICSEC. pp. 1–6. IEEE (2015)
2. Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., Rodriguez-Muro, M., Xiao, G.: Ontop: Answering sparql queries over relational databases. Semantic Web **8**(3), 471–487 (2017)
3. Espírito Santo, J.M.d., Medeiros, C.B.: Semantic interoperability of clinical data. In: The Int. Conf. on Data Integration in the Life Sciences (DILS). pp. 29–37. Springer International Publishing (2017)
4. Haendel, M.A., Chute, C.G., Robinson, P.N.: Classification, ontology, and precision medicine. New England Journal of Medicine **379**(15), 1452–1462 (2018)
5. Hodge, G.: Systems of knowledge organization for digital libraries : beyond traditional authority files. Washington, D.C. : Digital Library Federation (2000)
6. IHME: Global burden of disease study 2016 (gbd 2016) causes of death and nonfatal causes mapped to icd codes. http://ghdx.healthdata.org/record/ihme-data/gbd-2016-cause-icd-code-mappings, [accessed in Apr 1, 2019]
7. Lelong, R., Cabot, C., Soualmia, L.F., Darmoni, S.: Semantic search engine to query into electronic health records with a multiple-layer query language. In: Proc. of ACM SIGIR workshop on MedIR (2016)
8. Lim, L., Wang, H., Wang, M.: Semantic queries by example. In: Proc. of the ACM Int. Conf. on Extending Database Technology - EDBT. pp. 347–358 (2013)
9. Macário, C.G., Sousa, S.R., Medeiros, C.B.: Annotating Geospatial Data based on its Semantics. In: 17th ACM SIGSPATIAL Conference (2009)
10. Munir, K., Odeh, M., McClatchey, R.: Ontology assisted query reformulation using the semantic and assertion capabilities of owl-dl ontologies. In: Proc. of the ACM Int. Symposium on Database Engineering & Applications. pp. 81–90 (2008)
11. Sonntag, D., Möller, M.: Unifying semantic annotation and querying in biomedical image repositories. In: Proc. of KMIS. pp. 89–94 (2009)
12. Tiede, D., Baraldi, A., Sudmanns, M., Belgiu, M., Lang, S.: Architecture and prototypical implementation of a semantic querying system for big earth observation image bases. European journal of remote sensing **50**(1), 452–463 (2017)
13. Vilar, B.S.C.M.: Processamento de Consultas Baseado em Ontologias para Sistemas de Biodiversidade. Master's thesis, Instituto de Computação - Unicamp (2009)
14. Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A.: Bioportal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic acids research **39**(suppl_2), W541–W545 (2011)
15. Yunzhi, C., Huijuan, L., Shapiro, L., Travillian, R.S., Lanjuan, L.: An approach to semantic query expansion system based on hepatitis ontology. Journal of Biological Research-Thessaloniki **23**(1), 11 (2016)
16. Zhao, Y., Fesharaki, N.J., Li, X., Patrick, T.B., Luo, J.: Semantic-enhanced query expansion system for retrieving medical image notes. Journal of medical systems **42**, 1–11 (2018)
17. Zheng, S., Wang, F., Lu, J.: Enabling ontology based semantic queries in biomedical database systems. International journal of semantic computing **8**(01), 67–83 (2014)