Bridging the Gap between bdME and OntoME

Ricardo Giuliani Martini
Algoritmi Research Centre
Department of Informatics
University of Minho
Gualtar - 4710-057, Braga, Portugal
Email: rgm@algoritmi.uminho.pt

Pedro Rangel Henriques
Algoritmi Research Centre
Department of Informatics
University of Minho
Gualtar - 4710-057, Braga, Portugal
Email: prh@di.uminho.pt

***Abstract*—The Semantic Web aims at building a Web where data is enriched with meaningful annotations. In other words, data is semantically organized in such a way that both human and machine can understand and query it, aiming at the creation of dynamic Web pages. Ontologies, as a keystone of the Semantic Web, have gained an ample acceptance as an information model, which can be used for several purposes, such as information retrieval in the Web. However, data is normally stored in databases, which present various problems in the Semantic Web context, because data is not semantically annotated. Aiming at retrieving rich results in the sense of meaning, several ways of relating databases with ontologies have emerged. This paper presents a mapping – with the aid of a framework called Ontop – as a solution for the communication problem between the relational database of the Emigration Museum of Fafe (EMF) and the ontology of the Emigration Museum (OntoME), which describes the Cultural Heritage domain. This mapping will be used to realize the CaVa architecture, aiming at the creation of dynamic Web pages as virtual Learning Spaces. Real examples of the mapping process are presented.**

I. Introduction

Focusing on the creation of dynamic Web pages with rich content as virtual Learning Spaces (LS) about Cultural Heritage supported by documents, we came across the situation of dealing with the Semantic Web in order to facilitate the search, sharing, and combination of information related to a certain domain.

Taking into account the description of rich content sharable through the Web, ontologies play a crucial role, because they define the concepts that characterize the knowledge domain and their relationships allowing for a semantic manipulation of data. In this way, ontologies support a semantic search using annotations understandable by humans or machines (more details in [1]). To allow a sustainable growth of an ontology, efficient persistent storage of ontology concepts and data is essential [1].

The problem that makes this conceptualization difficult is that most data in the Web is stored in relational databases, due to their acceptance based on their adaptability, effectiveness, and performance for representing and managing data [2]. This means that to retrieve the information stored in those databases with the maximum possible wealth, it is necessary to relate databases with ontologies, the backbone of the Semantic Web.

Seeking the relationship between databases and ontologies, we have developed a mapping between them. Strictly speaking, this mapping serves to assign each concept and relation of the ontology to the data stored in the database for future exploitation.

To explain the mappings and instantiate the CaVa architecture (explained in Section II), in this work we use as a case study the Emigration phenomena in Portugal – more specifically the emigration movement from Fafe (a city in the north of Portugal) towards Europe or Brasil dated from 1960 to 1970.

To have access to those emigration documents, we have worked in collaboration with the Municipal Archive of Fafe, which holds fonds related to the emigration domain like passport application forms, ship routes, biographies, almanacs, among others[1]. At this moment, for this case study, we are interested in the passport application forms only. The structure, the information contained and how the passport application forms were stored in a digital format (bdME) were already published in [3].

So, with the goal of creating Web pages with rich content, we felt the need to elevate the emigration documents data to a more conceptual level. Thus, we did a search for ontologies that describe the Cultural Heritage domain and we have found the International Committee for Documentation – Conceptual Reference Model (CIDOC-CRM[2]), which is a formal ontology intended to facilitate the integration, mediation and interchange of heterogeneous Cultural Heritage information [4]. A study about the CIDOC Conceptual Reference Model and first drafts of the emigration ontology were presented in [5]. After that, for a successful development of the ontology, a "Reduced CRM-Compatible Form" ontology for the virtual Emigration Museum (OntoME) was built and published in [6].

The rest of the paper is structured as follows. Section II presents CaVa, an architecture for the automatic creation of virtual Learning Spaces. Section III details the knowledge management, introducing the database (bdME) and the OntoME ontology. Section IV describes Ontology-Based Data Access (OBDA). The mapping between bdME and OntoME in the context of the CaVa project is explained in Section V. Finally, Section VI presents our final thoughts.

[1] Although not created as works of art, those documents contain information that after processed can be considered as historic cultural material and in that sense can integrate in a museum collection.

[2] http://www.cidoc-crm.org/

II. The CaVa architecture

Regarding the creation of virtual Learning Spaces based on ontologies, the CaVa architecture[7] (see Figure 1) is presented in this section. To achieve the main goal of this proposal, an ontology should be used to describe the institutional information repository.

A Domain Specific Language (DSL) specification – specification of the virtual environment – shall be done. The LS specification shall delineate which concepts and how they should be placed in the final virtual LS.

The CaVa architecture aims at covering any specific domain, be it a museum, a library, an archive or a school. In other words, it was not planned for a specific domain.

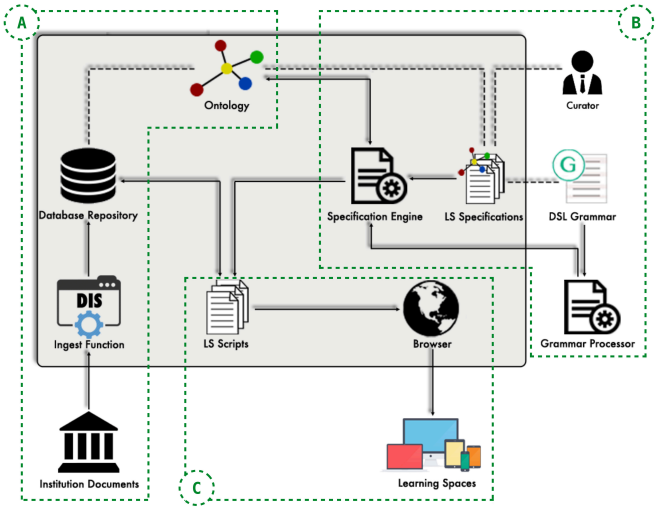

Fig. 1: The CaVa architecture

The CaVa architecture consists of three modules, which are depicted in Figure 1.

Module A comprises the *Institution Documents*, which holds the real documents; the *Database Repository*, which contains the documents data in an electronic storage format; a *Document Ingestion System* (DIS), to transfer the physical documents information from the *Institution Documents* to the *Database Repository*, and to aid in data management. Moreover, Module A includes the main *Ontology*, to describe, as a conceptual layer, the concepts and relations of the intended domain [3][6].

Module B is constituted by two important pieces: the *Specification Engine* and *LS Specifications* written in a Domain Specific Language designed particularly for this project. Observe that the *DSL grammar* is transformed in the *Specification Engine* in an automatic way by a compiler generator.

Module C includes the script files (*LS Scripts*) automatically generated by the *Specification Engine*; the *Browser*, that interprets and renders the final virtual environment. The output, also called virtual *Learning Space*, is the last component of Module C.

III. Knowledge Handling

This section presents the scenario of the *Emigration phenomena in Portugal*, showing the bdME database and the OntoME ontology. An Example is given to illustrate the structure of bdME and OntoME.

A. bdME and OntoME

bdME is a database that holds the data about the *Emigration phenomena in Portugal*, more specifically the passport application forms about the emigration movement of Fafe dated from 1960 to 1970. The database (bdME) contains more than 6000 documents. Each document has more than 80 data items (some fields are atomic, other lists, and some are compulsory while others are empty). The bdME schema consists of 16 tables. The respective database was built in MySQL Relational Database Management System (RDBMS).

This large set of data contains a high potential to describe each individual integrated in the society of his/her epoch. Besides, it also provides knowledge about the society in a specific context of the country's history [3].

An example of the data contained in bdME is shown in Table I. This data excerpt will be used as a running example to demonstrate the mapping in Section V.

TABLE I: *"identificacaoEmigrante", "filiacao", and "localidade" tables*

identificacaoEmigrante

idEmigrante	**nome**	**dtNasc**	***idFiliacao***	***idNaturalidade***
713204	Aníbal de Castro	1926-10-30	265	30712
720807	Lucinda Ribeiro	1935-07-29	37	30728
2665155	Manuel Vaz	1924-03-26	3	30717

filiacao

idFiliacao	**nomePai**	**nomeMae**
3	Júlio Vaz	Virgínia Delgado
37	Júlio Ribeiro	Maria Nogueira
265	–	Rosa de Castro

localidade

idLocalidade	**freguesia**	**concelho**	**distrito**
30712	Fornelos	Fafe	Braga
30717	Monte	Fafe	Braga
30728	São Gens	Fafe	Braga

Table I shows three tables of bdME that describe the data about the emigrant's identification (*identificacaoEmigrante*). The *identificacaoEmigrante* table comprises the identification number (idEmigrante), name (nome), birthdate (dtNasc), and two foreign keys that are related to the filiation (*filiacao* table) referenced by *idFiliacao* and birthplace (*localidade* table) referenced by *idNaturalidade*.

So, for instance, the emigrant identified by *2665155* has name *Manuel Vaz*, his birthdate is *1924-03-26* and has parents (filiation) identified by *3* – which corresponds to father *Júlio Vaz* and mother *Virgínia Delgado* from "filiacao" table – and his birthplace is referenced by *30717* – which corresponds to *Monte* parish, *Fafe* council, and *Braga* district, that together form a geographical location.

Looking at Table I, the data can be understood and we can infer knowledge from that data, but normally, the user

has no access to the schema and data of the databases. According to [2], information contained in databases cannot be semantically annotated. The authors describe this as "on one hand, the content of these databases is only shown when a query is performed in the database, and on the other hand, the semantic description of the database is represented using its schema, often unavailable or even useless because it can not be exploited depending of the format chosen to represent it". So, this means that without both the data and schema, it is difficult to extract some information from relational databases.

Taking the data of Table I as an example, we can see *idFiliacao* in *identificacaoEmigrante* table with values like "3", "37", and "265". Those values are references to *filiacao* table. The value "3", for instance, is a reference to a tuple that has two other values: "Júlio Vaz" and "Virgínia Delgado". Those data items are sequences of characters and can mean anything. Can we say that two values are the father and mother of the emigrant identified by *2665155* or his children?

Looking now to the *localidade* table, we can see an identifier (*idLocalidade*) and three other fields that correspond to parish (*freguesia*), council (*concelho*), and district (*distrito*). Those data items, together, can be understood as a locality (geographical place) and, indeed, they are, but do those places correspond to the birthplace or to the emigrant's home address? What do they really mean?

As said before, it is difficult to know the precise meaning of each data item without the association of the related concept chosen in a known vocabulary. It is only data, not information i.e., without a context, we cannot be sure what the data means.

However, having access to the conceptual layer, the user can query the database through the ontology concepts (known vocabulary) and the system should reason and translate the query into appropriate database questions [8]. An example of good questions to answer the previously mentioned issues are: who are the parents of the emigrant identified by *2665155*? What is the birthdate of the emigrant identified by *2665155*? These kind of questions involve concepts (parents, emigrant, and birth) that are well known, because we share a common vocabulary related to a specific context, i.e., the ontology vocabulary.

Based on the idea here discussed, we developed an ontology called OntoME, a CIDOC-Conceptual Reference Model, that describes the Cultural Heritage domain, particularly the emigration one. The blue rectangles of Figure 2 depict an excerpt of the ontology built to comprise this domain. Section V describes the mapping between the database and the ontology here considered.

IV. Ontology-Based Data Access (OBDA)

As said before, we aim to query the data layer (sources) through a conceptual layer. Ontology-Based Data Access (OBDA) is a paradigm that provides semantic access to databases by means of ontologies [9]. So, the goal of OBDA is to access and use data through an ontology.

OBDA has various attractive features, many of them have been already proved effective in managing complex information systems [10].

In OBDA, an abstract layer exists as an ontology, which defines a shared vocabulary of a specific domain. Besides, OBDA hides the database repository structure, and with this, it can enrich incomplete data with background knowledge [11].

OBDA is based on a three-level architecture established by an ontology (the main component and a formal description of the domain of interest), data sources, and the mappings linking the first two ones [10].

Based on the OBDA architecture, a lot of systems like Ontop[3], D2RQ[4], MASTRO[5], Ultrawrap[6], Morph-RDB[7], among others, have emerged in order to allow the users to query the sources from a conceptual view. In other words, those systems arose in order to facilitate the users life, so the users do not need to know anything about the data sources structure, they only need to know the domain in question. This is due to the whole process of translating user queries into the data vocabulary and assigning the responsibility of the actual query evaluation to the data sources being done by OBDA systems.

As previously mentioned, Ontop is used in this work to aid in the mappings between bdME and OntoME.

Ontop is an open-source project, developed at the Free University of Bozen-Bolzano, Italy. The Ontop system interprets SPARQL queries by rewriting them into Structured Query Language (SQL) queries over the database [12]. In this work Ontop is used as a Java API called OWLAPI[8].

V. bdME2OntoME Mapping

So, to provide the curator with access to the factual data in the database through the lens of the ontology, we have developed a mapping between the bdME database and the OntoME ontology using the Ontop framework. In that way the user can access the source through the OntoME's vocabulary.

The Ontop mapping needs to connect the classes and properties (datatype and object properties) of the ontology with views (SQL) over the repository's data via a specification, as can be seen in Listing I.

Listing I presents the mapping declarations for the bdME "identificacaoEmigrante", "filiacao", and "localidade" tables previously shown in Table I. The specification of the mapping should contain one or more mapping axioms[9] that intend to transform the repository's data into a set of RDF triples. This specification consists of three fields:

- *mappingId*: a string that identifies the mapping axiom;
- *target*: an ontology triple template (subject, predicate, object) which references column names used in the database

[3]available at: http://ontop.inf.unibz.it
[4]available at: http://d2rq.org/
[5]available at: http://www.dis.uniroma1.it/~mastro/?q=node/2
[6]available at: https://capsenta.com/ultrawrap/
[7]available at: https://github.com/oeg-upm/morph-rdb
[8]available at: http://owlapi.sourceforge.net/
[9]More about the mapping axioms: https://github.com/ontop/ontop/wiki/ontopOBDAModel#Mapping_axioms

(source) query through placeholders (terms between curly brackets);

- *source*: a SQL query over the database, which should contain the column names used in the target's placeholders.

LISTING I: *Mapping of bdME "identificacaoEmigrante", "filiacao", and "localidade" tables*

OBDA model

```
.....
[MappingDeclaration] @collection [[
mappingId   Emigrante
target      :Emigrante#{idEmigrante} a :E21.1_Emigrante ;
            :P131_is_identified_by :nomeEmigrante/{idEmigrante} ;
            :P152_has_parent :Filiacao#{idFiliacao} ;
            :P98i_was_born :nascimentoEmigrante#{idEmigrante} .
source      SELECT idEmigrante, idFiliacao FROM
            identificacaoEmigrante

mappingId   Emigrante Appellation
target      :nomeEmigrante/{idEmigrante} a :E82_Actor_Appellation ;
            :P3_has_note {nome} ; :P2_has_type ''Emigrante'' .
source      SELECT idEmigrante, nome FROM identificacaoEmigrante

.....
]]
```

In addition to the mapping declaration of OBDA model, there are two other sections for declaring the prefixes (`[PrefixDeclaration]`) and the source (`[SourceDeclaration]`) needed by the mapping statement. This should be done before the `[MappingDeclaration]` section of Listing I, represented here by the "....." sign.

To better illustrate the mapping axioms, Figure 2 presents the same mapping declarations of Listing I, but in a graphical format, which aids in the interpretation of the mapping rules, because we can "navigate" over the concepts and relations until we reach the placeholders.

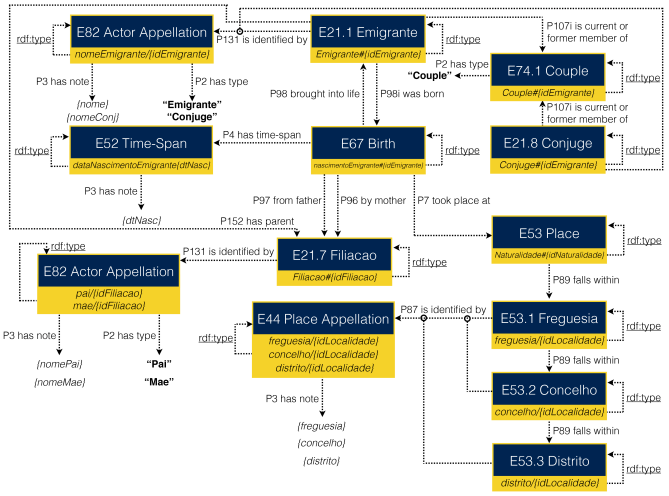

Fig. 2: Example describing the *"Emigrante"*, *"Filiacao"*, and *"Place"* concepts mapping

Taking the second mapping (*mappingId* Emigrante Appellation) of Listing I as an example, we can see two kinds of placeholders: (a) literal value – `{nome}`; and (b) a Uniform Resource Identifier (URI) template – `:nomeEmigrante/{idEmigrante}`, where the colon (":") sign corresponds to the CIDOC-CRM ontology's prefix (in this case `http://erlangen-crm.org/current/`). The first kind of placeholder is used to generate literal values, while the second one serves to construct object URIs from the repository's data.

The goal of the mapping axioms, as already mentioned, is to convert the sources data into RDF triples. To be more specific, those mapping axioms generate a set of RDF triples for each returned result of the query over the database (represented in Listing I by source).

So, using Table I data (only the *identificacaoEmigrante* database table) and the second mapping of Listing I as example, the result of the mapping is the following set of nine RDF triples (presented in Turtle[10] syntax):

Generated RDF triples (represented in Turtle syntax)

```
:nomeEmigrante/713204    rdf:type        :E82_Actor_Appellation ;
                         :P3_has_note    ''Aníbal de Castro'' ;
                         :P2_has_type    ''Emigrante'' .

:nomeEmigrante/720807    rdf:type        :E82_Actor_Appellation ;
                         :P3_has_note    ''Lucinda Ribeiro'' ;
                         :P2_has_type    ''Emigrante'' .

:nomeEmigrante/2665155   rdf:type        :E82_Actor_Appellation ;
                         :P3_has_note    ''Manuel Vaz'' ;
                         :P2_has_type    ''Emigrante'' .
```

Note the substitution of the placeholders `{nome}` as literal value and `{idEmigrante}` in the URI template `:nomeEmigrante/{idEmigrante}` by the values of the *identificacaoEmigrante* table of Table I.

After the RDF triples are generated, the dataset can be queried by SPARQL, i.e., using the vocabulary specified in OntoME. As previously mentioned, we have used OWLAPI to query the sources through the ontology.

To exemplify a query over the generated dataset, we will show a query that retrieves all the emigrants' names and their respective birthdate, ordered by the name of the emigrant.

SPARQL query example

```
SELECT distinct ?name ?dt
WHERE {
   ?idE a :E21.1_Emigrante ;
        :P131_is_identified_by ?emAppellation ;
        :P98i_was_born ?birth .
   ?emAppellation :P3_has_note ?name ;
                  :P2_has_type ''Emigrante'' .
   ?birth :P4_has_time-span ?dtBirth .
   ?dtBirth :P3_has_note ?dt .
} ORDER BY (?name)
```

This SPARQL query retrieves all the emigrant's names (`?name`) and the birthdate (`?dt`) related to them from the

[10]https://www.w3.org/TR/turtle/

bdME database. To achieve the right result, the basic graph pattern[11] is specified in the `WHERE` clause to match with the data graph. In this case, the `?idE` variable should be an instance of the `E21.1_Emigrante` concept. It must also be identified by an appellation (`:P131_is_identified_by ?emAppellation`) and be related to a birth event (`:P98i_was_born ?birth`).

After that, to match the data graph, the `?emAppellation` should have, as a note, the name (`?name`) listed in the `SELECT` clause. Moreover, the variable of the birth concept (`?birth`) should have a time-span here mentioned as the variable `?dtBirth` (`:P4_has_time-span ?dtBirth`), which, in turn, needs to have, as a note, the birthdate (`?dt`) listed in the `SELECT` clause. To get an ordered list of emigrants by name, the `ORDER BY` clause is necessary.

Observe that the mapping declaration pattern

```
:nomeEmigrante/{idEmigrante} a :E82_Actor_Appellation
```

for the `?emAppellation` variable in the SPARQL query was not needed, because Ontop can infer it from

```
?idE :P131_is_identified_by ?emAppellation
```

since the range of `:P131_is_identified_by` is `:E82_Actor_Appellation`, as specified in the mapping declaration.

An important thing to note is that the vocabulary of the query is driven by the domain of interest and it is independent of the database. This task paved the way to allow the end-user to query emigrant's repository not from a database perspective, but driven by its conceptual model.

VI. Conclusion

We have analysed the ontology-based data access issue from the perspective of maintaining the conceptual layer and the data sources separate and independent. The solution provided in this paper is based on the adoption of the Ontop framework that links the conceptual schema to the factual database.

We have implemented our solution on top of the CaVa architecture, which aims at the automatic creation of virtual Learning Spaces. The mapping here discussed in the context of the CaVa project enables the specification of virtual learning environments in a concept-based perspective (conceptual layer), giving the curator the opportunity of specifying virtual exhibition rooms without the need of knowing the sources' structure.

After a deep reflection about the mapping concept and creation we built a tool, called CaVa^{MG} (mapping generator), to aid in the creation of the mapping axioms. The tool allows to choose the database and the ontology and automatically lists all the elements available to concretize the mapping.

The next step is to test intensively the proposal in real case studies to understand the actual applicability of the methodological approach and access the performance attained.

[11] Basic graph patterns are sets of triple patterns. To learn more about it: https://www.w3.org/TR/rdf-sparql-query/#BasicGraphPatterns

Acknowledgment

This work has been supported by COMPETE: POCI-01-0145-FEDER-007043 and FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013. The work of Ricardo Giuliani Martini is supported by CNPq, grant 201772/2014-0.

References

[1] A. Gali, C. X. Chen, K. T. Claypool, and R. Uceda-Sosa, *Conceptual Modeling for Advanced Application Domains: ER 2004 Workshops CoMoGIS, CoMWIM, ECDM, CoMoA, DGOV, and eCOMO, Shanghai, China, November 8-12, 2004. Proceedings*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, ch. From Ontology to Relational Databases, pp. 278–289. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-30466-1_26

[2] C. Martinez-Cruz, I. J. Blanco, and M. A. Vila, "Ontologies versus Relational Databases: Are they so different? A comparison," *Artif. Intell. Rev.*, vol. 38, no. 4, pp. 271–290, Dec. 2012. [Online]. Available: http://dx.doi.org/10.1007/s10462-011-9251-9

[3] R. Martini, M. Guimarães, G. Librelotto, and P. Henriques, "Storing archival emigration documents to create virtual exhibition rooms," in *New Contributions in Information Systems and Technologies*, ser. Advances in Intelligent Systems and Computing, A. Rocha, A. M. Correia, S. Costanzo, and L. P. Reis, Eds. Springer International Publishing, 2015, vol. 353, pp. 403–409. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16486-1_40

[4] ICOM/CIDOC, "Definition of the cidoc conceptual reference model," ICOM/CIDOC, Tech. Rep., May 2015. [Online]. Available: http://www.cidoc-crm.org/docs/cidoc_crm_version_6.2.pdf

[5] R. Martini, "Formal Description and Automatic Generation of Learning Spaces based on Ontologies," Universidade do Minho, PhD pre-thesis, 2015.

[6] R. G. Martini, C. Araújo, G. R. Librelotto, and P. R. Henriques, "A Reduced CRM-Compatible Form Ontology for the Virtual Emigration Museum," in *New Advances in Information Systems and Technologies*, Á. Rocha, M. A. Correia, H. Adeli, P. L. Reis, and M. M. Teixeira, Eds. Cham: Springer International Publishing, 2016, pp. 401–410. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-31232-3_38

[7] R. G. Martini, G. R. Librelotto, and P. R. Henriques, "Formal Description and Automatic Generation of Learning Spaces based on Ontologies," in *20th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems, KES2016*. Elsevier, September 2016, to be published.

[8] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati, "Linking data to ontologies," in *Journal on Data Semantics X*, S. Spaccapietra, Ed. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 133–173.

[9] M. R. Kogalovsky, "Ontology-based data access systems," *Programming and Computer Software*, vol. 38, no. 4, pp. 167–182, 2012. [Online]. Available: http://dx.doi.org/10.1134/S0361768812040032

[10] M. Lenzerini, "Ontology-based data management," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: ACM, 2011, pp. 5–6. [Online]. Available: http://doi.acm.org/10.1145/2063576.2063582

[11] R. Kontchakov, M. Rezk, M. Rodríguez-Muro, G. Xiao, and M. Zakharyaschev, "Answering sparql queries over databases under owl 2 ql entailment regime," in *Proceedings of the 13th International Semantic Web Conference - Part I*, ser. ISWC '14. New York, NY, USA: Springer-Verlag New York, Inc., 2014, pp. 552–567. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-11964-9_35

[12] D. Calvanese, B. Cogrel, S. Komla-Ebri, D. Lanti, M. Rezk, and G. Xiao, "How to stay ontop of your data: Databases, ontologies and more," in *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers*, 2015, pp. 20–25. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-25639-9_4