Benchmarking Applied to Semantic Conceptual Models of Linked Financial Data

José Luis Sánchez-Cervantes[1], Lisbeth Rodríguez-Mazahua[1], Giner Alor-Hernández[1],
Cuauhtémoc Sánchez-Ramírez[1], Jorge Luis García-Alcaráz[2],
and Emilio Jimenez-Macias[3]

[1] Division of Research and Postgraduate Studies, Instituto Tecnológico de Orizaba
Av. Oriente 9, 852. Col. Emiliano Zapata, 94320, Orizaba, México
isc.jolu@gmail.com, {lrodriguez,galor,csanchez}@itorizaba.edu.mx
[2] Department of Industrial Engineering, Universidad Autónoma de Ciudad Juárez
Av. del Charro, 450 Norte. Col. Partido Romero, 32310, Ciudad Juárez, México
jorge.garcia@uacj.mx
[3] Department of Electrical Engineering, University of La Rioja,
C/Luis de Ulloa, 20, 26004 Logroño, La Rioja, Spain
emilio.jimenez@unirioja.es

Abstract. Semantic modeling plays a central role in knowledge-based systems where information sharing and integration is a primary objective. Ontology and metadata description languages such as OWL (Web Ontology Language) and RDF(S) (Resource Description Framework Schema) are commonly the most used for representing semantic models and data. The graph-like structure adopted for semantic metadata representation allows simple and expressive queries by using SPARQL-based subgraph matching. While performance of such knowledge-based systems depends on multiple factors, in this work we present a mechanism to properly choice a semantic modeling pattern in order to significantly reduce the data query execution time. Based on this understanding, this work proposes a comparative analysis of different conceptual modeling approaches on the basis of financial domain. In order to show the efficiency/accuracy of our approach, an evaluation of SPARQL-based queries was performed against different modeled datasets.

Keywords: Conceptual modeling · Linked Data · Performance · Semantics; SPARQL

1 Introduction

Nowadays the Financial domain is a source of a great amount of data, as enterprises periodically publish information relative to their financial statements. However, there are multiple ways for representing this information. In financial environments, finding the right information at the right time is the key issue for decision-making process [1]. From this perspective, the importance of performance is twofold. On the one hand, finding information in a fast way could be critical for making an important decision

I. Ciuciu et al. (Eds.): OTM 2015 Workshops, LNCS 9416, pp. 289–298, 2015.
DOI: 10.1007/978-3-319-26138-6_32

and, on the other hand, due to the great volume of information, it is necessary to optimize the process. For this reason, an appropriate representation model could provide a common way for efficiently representing and retrieving financial information. This work is based on the hypothesis that the use of the appropriate semantic modeling pattern might reduce the data retrieval time through the SPARQL-based queries execution, and therefore the process of finding data and decision-making process can be more efficient. To confirm this hypothesis, a process for the extraction and processing of XBRL (eXtensible Business Reporting Language) financial statements published in the EDGAR (Electronic Data Gathering, Analysis, and Retrieval system)[1] repository using semantic technologies, such as RDF(S), OWL and SPARQL, was performed, with the aim of generating a financial knowledge base inspired on Linked Data principles [2] that is conformed of two different graphs assigned at Mixed and Entity-Attribute-Value (EAV) semantic models. Through these semantic models, we have designed and run a set of SPARQL-based queries with the aim of identifying which of these semantic models provides the acquisition of financial data faster.

This paper is organized as follows: Section 2 summarizes the Literature review; Section 3 presents the financial taxonomy and the two models (EAV and Mixed) used as basis of this research; Section 4 describes the experiment set up and Section 5 presents and discusses the obtained results; finally Section 6 presents conclusions and future work.

2 Literature Review

In the literature, there are many initiatives related with applying benchmark testing on RDF (Resource Description Framework) datasets corresponding to several domains. Some of these initiatives have obtained interesting results, which are briefly described below. Fundulaki et al. [3] presented the Linked Data Benchmark Council (LDBC) project with the aim of providing a solution to the following problems: a) the lack of a comprehensive suite of benchmarks that encourage the advancement of technology by providing both academia and industry with clear targets for performance and functionality; and b) the need for an independent authority for developing benchmarks and verifying the results of RDF engines. The solution to these problems was timely and urgent because non-relational data management is emerging as a critical need for the new data economy based on large, distributed, heterogeneous, and complexly structured datasets. Our proposal intends to contribute providing a benchmark to measure the time for information retrieval from the comparison of two models for semantic representation of financial data.

The Berlin SPARQL Benchmark (BSBM) for comparing the performance of several semantic systems, such as native RDF stores, systems that map relational databases into RDF, and SPARQL wrappers around other kinds of data sources across architectures, was presented by Bizer and Schultz [4]. FedBench, a comprehensive benchmark suite for testing and analyzing both the efficiency and effectiveness of federated query processing on semantic data was presented by Schmidt et al. [5].

[1] https://www.sec.gov/edgar/searchedgar/companysearch.html

An evaluation of FedBench, which is considered as the most comprehensive SPARQL testbed up to now, was presented by Montoya et al. [6]. The creation of a generic procedure SPARQL benchmark applied to the DBpedia base knowledge was proposed by Morsey et al. [7]. SRBench, a general-purpose benchmark primarily designed for streaming RDF/SPARQL engines, completely based on real-world data sets from the Linked Open Data cloud was introduced by Zhang et al. [8]. A benchmark for comparing the expressivity as well as the runtime performance of data translation systems, trough the design of LODIB (Linked Open Data Integration Benchmark) was presented in the work of Rivero et al. [9]. Bail et al. [10] presented FishMark, a Linked Data application benchmark to compare the performance of the native MySQL application, the Virtuoso RDF triple store, and the Quest OBDA system on a fishbase.org like application.

Unlike the initiatives [4-10], in this work several datasets are not compared. In our proposal, we have established a comparison between the EAV and Mixed models to represent financial information. Such comparison involves the execution of a set of SPARQL-based queries in order to measure the runtime of data retrieval in both models.

A classification methodology for federated SPARQL queries and a heuristic called SPLODGE for automatic generation of benchmark queries that is based on this methodology and takes into account the number of sources to be queried and several complexity parameters were presented by Görlitz et al. [11]. The RDF benchmark to model a large scale electronic publishing scenario was presented by Tarasova and Marx [12]. Unlike these initiatives, in our work, we propose a comparative analysis of two different conceptual modeling approaches on the basis of financial domain. The first contribution of the work presented by Aluç et al. [13] is an in-depth experimental analysis which shows that existing SPARQL benchmarks are not suitable for testing systems for diverse queries and varied workloads. To address these shortcomings, their second contribution is the Waterloo SPARQL Diversity Test Suite (WatDiv) that provides stress-testing tools for RDF data management systems. Our contributions are 1) Propose two semantic models inspired in Linked Data principles [2] in order to publish financial information from multiple sources; 2) Provide a benchmark that allows the definition of which financial Linked Data model presented is the most appropriate to publish, search and calculate financial information.

Some of previously works described have obtained outstanding results by applying benchmark tools over several datasets. A key challenge for the semantic Web is to acquire the capability to effectively query large knowledge bases. From this perspective, unlike these works, we describe two Semantic data models (EAV and Mixed models) with their respective benchmarking in order to compare their performance for data retrieval in a financial data context.

3 Description of Semantic Data Models

The overall objective of semantic data models is to capture more meaning of data by integrating relational concepts with more powerful abstraction concepts known from the Artificial Intelligence field in order to facilitate the representation of real world situations [14], [15]. For benchmark purposes, we evaluated two data models, the

EAV model and the Mixed model. Such models include public companies' financial statements reporting such as balance sheets, cash flow and income statements. This is an entry point for the general evaluation of the modeling techniques oriented at query performance and data retrieval. Each model semantically represents the interaction between classes and subclasses that integrate it, basing on a simplified financial taxonomy, generated from published Balance sheets under the US-GAAP principles, through the EDGAR repository.

3.1 Entity Attribute Value Model

The Entity-Attribute-Value (EAV) approach is popular for modeling highly heterogeneous data by using a relatively simple physical database schema (in database literature, alternative terms for entity and attribute are object and parameter, respectively) [16].

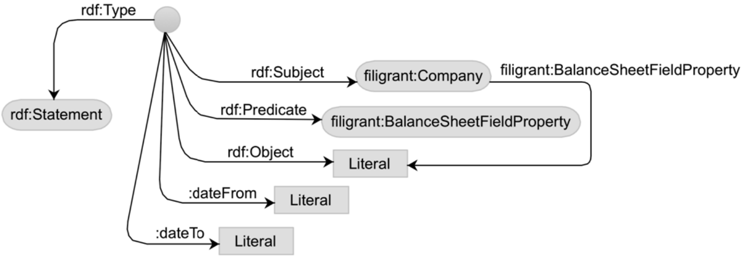

Fig. 1. Entity-Attribute-Value (EAV) Financial data model with reification

As shown in the Fig. 1, the EAV data model representation is as follows: *Entity* is the company name; *Attribute* corresponds to the financial ratio and; *Value* is the value assigned to financial ratio. Furthermore, this model uses the reification [17], [18] in order to attaching as properties the period (start and end dates) of publication of the balance sheets. Unlike the Mixed semantic model, the financial data transformation to RDF notation, following the EAV data model, uses the ratios of the taxonomy as properties (Attributes).

3.2 Mixed Model

The Mixed model (see Fig 2.) developed for the analysis of its performance in the data retrieval through the execution of SPARQL-based queries, is sustained in the EAV model, and in the canonical data model, also named Common Data Model (CDM), that allows defining the entities relevant for a specific domain, including their attributes, associations, and their semantics [19].

In the Mixed model, the data transformation to RDF notation uses the financial ratios specified in the simplified taxonomy as a class hierarchy, representing the inherent nature and characteristics of the financial data, its structure and how the different parts (classes, subclasses and values) are related to each other. The aim of this transformation is to provide a normalized model that is adjusted in a "natural way" in the simplified financial taxonomy, but keeping the features of the EAV model.

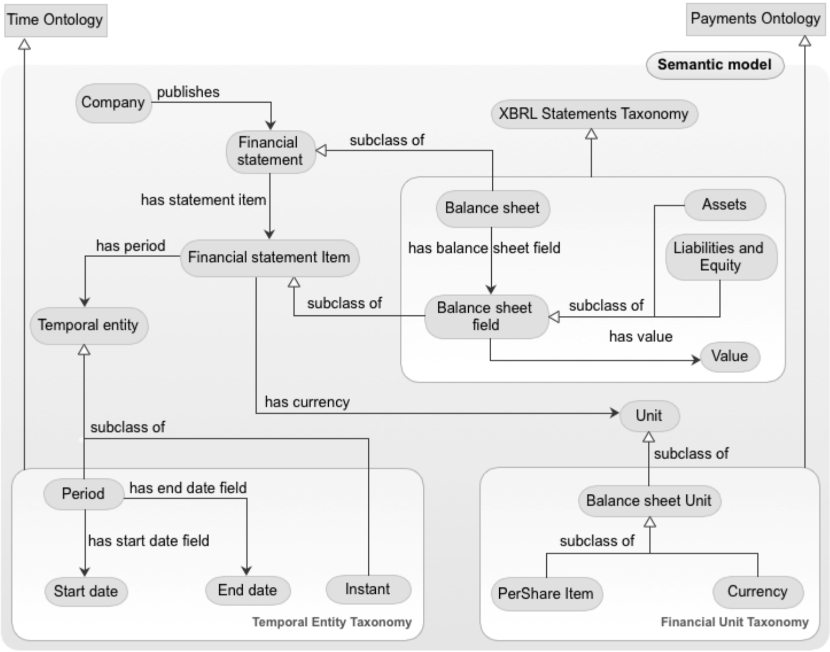

Fig. 2. Mixed Financial data model

4 Benchmark Experiments Design

We have provided two financial data collections; each one consists of a number of interlinked datasets. It is important to underline that we do not assess repositories per se nor we analyze them in terms of query complexity (such as Berlin Benchmark [4] and DBPedia SPARQL Benchmark [7]), but we focus on how semantic data modeling strategies can affect the performance of data retrieval. The experiment starts with the creation of semantic graphs of each semantic model for representing the same financial information. The data is obtained in the information extraction process from financial fillings corresponding to 830,321 XBRL files published by US companies in the EDGAR system. We performed the conceptual modeling part and we created two semantic schemas that they were later used to represent XBRL 10-Q reports data. As a result, we obtained two distinct RDF data graphs that represent the same information as the original data source, but structured after different semantic models. The graphs were loaded to separate semantic repositories, and some series of prepared SPARQL-based queries typical of the financial domain (see Table 1) were issued to measure the execution time that each model takes for data retrieval.

Table 1. SPARQL-based queries for benchmark experiments

SPARQL Query	Description
Q-1	It retrieves all the information of the first 500,000 records in the dataset.
Q-2	It gets a list with the company name and Central Index Key (CIK) registered in the dataset.
Q-3	It gets the financial concepts related to a company (e.g. Apple Inc.), indicating the document period end date for each financial concept.
Q-4	From Google, Microsoft and Yahoo companies, it retrieves information from the ratios of their balance sheets with their respective values published the following dates is between 01/01/2010 and 31/12/2012.
Q-5	It retrieves information of companies whose their Goodwill Value is greater than 10,000,000,000dlls and their document fiscal year focus is 2013.
Q-6	It calculates the Acid test for ABTECH HOLDINGS, INC. It is based on the fiscal focus indicator and the fiscal year focus value. Acid Test is an accounting ratio that indicates the liquidity or solvency of a company in the short term [20].
Q-7	It calculates the Day Time Interval Measurement for ABTECH HOLDINGS, INC. It is based on the fiscal focus indicator and the fiscal year focus value. The Day Time Interval Measurement calculation allows getting the number of days in which a company can continue operating, if for some reason, the company stops its daily activities [20].

The experiments have been carried out under the following technological capabilities: a computer of 64 bits with Operating System Windows Server 2008 R2 Standard, Service Pack 1,8 GB of RAM, a processor AMD Phenom(tm) II X6 1090 3.20GHz and Virtuoso Open-Source Edition (version 7.2.1) as support platform to the RDF triplets.

We decided to use Virtuoso Open-Source because we believe that it is the platform for management, access and integration of Linked Data more convenient to perform our experiments. Our decision is based on the results of benchmark tests for the execution of SPARQL-based queries performed by other authors, such as Bizer and Schultz [4] and Morsey et al. [7], which compared various systems for managing data based on Linked Data, and their obtained results indicated that Virtuoso was the fastest. The results obtained after the execution of the set of SPARQL-based queries for both models are described in the next section.

5 Results

First, we have considered to measure the loading time of the triples in Virtuoso Open-Source for both models. The mixed model obtained a loading time of 3,9 hours with 4,76 GB of files with triples, while the EAV model obtained a loading time of 5,18 hours with 5,84 GB of files with triples. The loading process of the triples in the dataset generated a total of 138, 675, 457 triples, of which 89,977,851 correspond to the

EAV model and 48,697,606 triples to the graph of Mixed model. For both models, the set of SPARQL-based queries was executed five times with the purpose of finding the same information and to calculate the average runtime of the data retrieval. These SPARQL-based queries were executed through the iSQL tool of Virtuoso Open-Source, and the results were dumped into .txt files. The records stored in these files allow analyzing the data retrieved. The main metric used to compare the obtained results of the EAV and Mixed models is the runtime of data retrieval (measured in milliseconds/ms). These results are shown in the Table 2 and are available (.zip) on the following URL:

https://drive.google.com/file/d/0B1dT-T9E25tTUVJLc2hYeTFHQnM/view?usp=sharing

Table 2. Benchmark time of data retrieval for EAV and Mixed models

Time to data retrieval (ms)							
MODEL	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**
EAV	37737	920	748	3198	826	2870	550
	38095	858	733	3213	827	3447	560
	38048	890	764	3042	1404	3588	550
	35210	874	827	3183	780	3354	550
	37690	874	734	3214	780	3292	560
Mixed	47705	499	47	3261	827	952	484
	31808	1529	31	3276	624	967	515
	35990	1560	31	2121	889	936	468
	35990	1653	16	3276	687	936	499
	36083	1279	31	3183	827	952	483
Averages (ms)							
EAV	37356	883,2	761,2	3170	923,4	3310,2	554
Mixed	37515,2	1304	31,2	3023,4	770,8	948,6	489,8

The following program listing is an example of a SPARQL-based query executed for the Mixed model

```
SELECT DISTINCT ?companyName ?goodwillValue ?documentFiscalYearFocus
WHERE { ?s flgrant:EntityRegistrantName ?companyName .
    ?s flgrant:hasBalanceSheetField ?BalanceSheetField .
    ?s flgrant:DocumentFiscalYearFocus ?documentFiscalYearFocus .
    ?BalanceSheetField flgrant:hasMonetaryValue ?monetaryValue .
    ?monetaryValue flgrant:value ?goodwillValue .
    ?BalanceSheetField a ?bsclass .
    ?bsclass rdfs:label ?BalanceSheetFieldLabel .
    FILTER (?BalanceSheetFieldLabel = "Goodwill")
    FILTER (xsd:integer(?goodwillValue) >= 10000000000)
    FILTER (xsd:integer(?documentFiscalYearFocus)="2013"^^xsd:integer)};
```

The above SPARQL-based query is an example of obtaining the results of Q5 corresponding to the Mixed model. For simplicity, we skipped the prefixes in this example.

The initial experiments were performed to the EAV model and its average results favor to the Q2, Q3, Q5 and Q7 queries with values of less than 1000ms for each query. Q2, Q3 and Q5 are medium complexity queries that require retrieving data based on one or two search criteria. Q5 is slightly more complex because it requires searching those values corresponding to the Goodwill ratio with value of more than 1000,000,000dlls. In contrast, Q7 requires a series of calculations for the Daytime interval measurement based on a particular Document Fiscal Focus and a Focus Document Fiscal Year. Moreover, Q1 and Q4 are queries that require retrieving a considerable amount of data. The first one requires retrieving the general information of the first 500,000 records of all triples generated in this model, while the second one requires obtaining data within a date range. The times obtained for these queries are reasonable, considering that the subject of triples generated for this model serves as an index, which is the retrieval path for the desired data during the search process. However, the time obtained in Q7, compared to Q6, is very good, indicating that this model is useful for certain calculations.

If the results of each model are analyzed one by one in the Table 2, we can find two notable differences; the first one indicates that the average time for data retrieval in Q2 is higher in the Mixed model compared with the EAV model. However, it is not the same case for Q6. Other queries have certain similarities in both models, for example queries Q1 and Q4 exceed 1000ms, while Q3, Q5 and Q7 remain this value. The average times presented in Table 2 show that the Mixed model scored the best times for data retrieval in processed queries, with exception of Q1 and Q2. However, the difference between the two models does not exceed the 500ms. The overall average time for EAV model is 6708,285,714ms, and on the Mixed model is 6297,571,429ms, with a difference of 410,714,285ms. Therefore, we deduce that the Mixed model is the most optimal for the execution of queries processed in these experiments.

6 Conclusions and Future Work

Knowledge representation is the basis for sharing and knowledge reuse. In this way, Ontologies and Linked Data provide the structure and the tools for representing and sharing knowledge allowing information retrieval based on common vocabularies. These characteristics are especially relevant for the financial domain where the data sources are diverse and appropriate semantic models are necessary for representing and retrieving information. Based on this understanding, the calculation of rations in the financial domain is particularly relevant. However, performance issues must be taken into account in order to provide the right information at the right time. For this reason, in this paper two conceptual modeling approaches for the financial domain have been presented.

Both conceptual models are based on the simplified US-GAAP Balance Sheet Taxonomy. These models allow the representation of financial ratios as well as perform searches. Despite that the EAV model is optimal for directly browsing and finding information, the Mixed model proposed favors the financial calculations and the search of ratios. This model is especially relevant for the financial domain where information is usually result of calculations based on data not directly accessible. Thus, the results of the benchmark analysis of both approaches showed that the Mixed model is the most optimal for the execution of the SPARQL-based queries corresponding to the context of our work.

Based on the results of this study, future research will include the execution of more experiments with the processing of SPARQL-based queries, as well as the calculating of Student's T-distribution to corroborate statistically if the Mixed model continues acquiring the financial data faster. Furthermore, we pretend adding an extension of the Mixed Model in order to provide a Linked-Data based framework for representing the financial data of enterprises which publish their results in order to provide an efficient environment for sharing financial information. Such Linked Data approach will connect the financial data with other data sources in order to enrich the information and provide efficient added value services.

Acknowledgement. The authors are very grateful to National Technological of Mexico for supporting this work. Also, this research paper was sponsored by the National Council of Science and Technology (CONACYT), as well as by the Public Education Secretary (SEP) through PRODEP.

References

1. O'Riain, S., Curry, E., Harth, A.: XBRL and open data for global financial ecosystems: A linked data approach. Int. J. Account. Inf. Syst. **13**(2), 141–162 (2012)
2. Berners-Lee, T.: Linked Data - Design Issues. Linked Data (2009). http://www.w3.org/DesignIssues/LinkedData.html (Accessed October 08, 2013)
3. Fundulaki, I., Pey, J.L., Dominguez-Sal, D., Toma, I., Fensel, D., Bishop, B., Neumann, T., Erling, O., Neubauer, P., Groth, P., Van Harmelen, F., Boncz, P.: The Linked Data Benchmark Council (LDBC). Proc. First Eur. Data Forum **877**, 6–8 (2012)
4. Bizer, C., Schultz, A.: The Berlin SPARQL Benchmark. Int. J. Semant. Web Inf. Syst. **5**(2), 1–24 (2009)
5. Schmidt, M., Görlitz, O., Haase, P., Ladwig, G., Schwarte, A., Tran, T.: FedBench: a benchmark suite for federated semantic data query processing. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 585–600. Springer, Heidelberg (2011)
6. Montoya, G., Vidal, M.-E., Corcho, O., Ruckhaus, E., Buil-Aranda, C.: Benchmarking federated sparql query engines: are existing testbeds enough? In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part II. LNCS, vol. 7650, pp. 313–324. Springer, Heidelberg (2012)
7. Morsey, M., Lehmann, J., Auer, S., Ngonga Ngomo, A.-C.: DBpedia SPARQL benchmark – performance assessment with real queries on real data. In: Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E. (eds.) ISWC 2011, Part I. LNCS, vol. 7031, pp. 454–469. Springer, Heidelberg (2011)
8. Zhang, Y., Duc, P.M., Corcho, O., Calbimonte, J.-P.: SRBench: a streaming RDF/SPARQL benchmark. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 641–657. Springer, Heidelberg (2012)
9. Rivero, C.R., Schultz, A., Bizer, C., Ruiz, D.: Benchmarking the performance of linked data translation systems. In: Linked Data on the Web (LDOW 2012) workshop (2012)
10. Bail, S., Alkiviadous, S., Parsia, B., Workman, D., Van Harmelen, M., Concalves, R.S., Garilao, C.: FishMark: A linked data application benchmark (2012)

11. Görlitz, O., Thimm, M., Staab, S.: SPLODGE: systematic generation of SPARQL benchmark queries for linked open data. In: Cudré-Mauroux, P., Heflin, J., Sirin, E., Tudorache, T., Euzenat, J., Hauswirth, M., Parreira, J.X., Hendler, J., Schreiber, G., Bernstein, A., Blomqvist, E. (eds.) ISWC 2012, Part I. LNCS, vol. 7649, pp. 116–132. Springer, Heidelberg (2012)
12. Tarasova, T., Marx, M.: ParlBench: a SPARQL benchmark for electronic publishing applications. In: Cimiano, P., Fernández, M., Lopez, V., Schlobach, S., Völker, J. (eds.) ESWC 2013. LNCS, vol. 7955, pp. 5–21. Springer, Heidelberg (2013)
13. Aluç, G., Hartig, O., Özsu, M., Daudjee, K.: Diversified stress testing of rdf data management systems. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 197–212. Springer, Heidelberg (2014)
14. Klas, W., Schrefl, M.: Semantic data modelling. Metaclasses Their Appl. Data Model Tailoring Database Integr, 71–81 (1995)
15. NIST-FIPS: Integration definition for information modeling (IDEF0-IDEF1X) (1993)
16. Nadkarni, P.M., Marenco, L., Chen, R., Skoufos, E., Shepherd, G., Miller, P.: Organization of Heterogeneous Scientific Data Using the EAV/CR Representation. J. Am. Med. Informatics Assoc. **6**(6), 478–493 (1999)
17. Alexander, N., Ravada, S.: RDF object type and reification in the database. In: Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, p. 93 (2006)
18. Manola, F., Miller, E., McBride, B.: RDF primer. W3C Recomm. **10**, 1–107 (2004)
19. Dell, M., Dell, S.: Canonical Data Model Design Guidelines (2010)
20. Montero, J.M., Fernández-Aviles, G.: Enciclopedia de economía, finanzas y negocios. Editorial CISS (Grupo Wolters Kluwer), Madrid (2010)