Ontology as mapping of material world

Irina Leshcheva[1*], and *Dmitry* Leshchev[2]

[1] Saint Petersburg State University, the Institute "Graduate School of Management", 199004 Volkhovsky per. 3, Russian Federation
[2] Peter the Great Saint Petersburg Polytechnic University, Centre for Advanced Studies, 195251 Polytechnicheskaya st. 29, Russian Federation

Abstract. This paper reviews currently available approaches and methods of automated population and enrichment of ontologies or ontology-based knowledge bases by structured data stored in various heterogeneous sources. Advantages and disadvantages of each approach are pointed out. The results of the analysis allow concluding that the existing methods are not effective enough to solve practical problems. A new method suggested does not have any of the specified disadvantages. The suggested method allows integrating data from different types of sources and considering the distributed nature of data and the necessity of authentication during accessing network resources. The method also suggests a solution to the problem of integrated data conflict resolution so that it will reduce the complexity of populating and enriching ontologies using the collected data arrays, irrespective of the format in which data are stored or represented.

1 Introduction

The term "ontology" came from Philosophy where it means the theory of existence; the subdiscipline of Philosophy that studies the fundamental principles of existence, the most general entities and their categories [1]. In the end of the 20th century, the term "ontology" started to be used in the artificial intelligence, in particular, knowledge engineering [2].

There are many approaches to defining the term "ontology". Tom Gruber formulated one of the most well-known definitions of ontology: "Ontology is a specification of conceptualization" [3]. "Conceptualization" is understood as a rigorous description of the system of concepts, objects and other entities and relationships between them [4]. In other words, conceptualization is a simplified model of the world created for certain purposes using the systems approach. Accordingly, ontology may be considered as mapping of the material world to a certain structured verbal or symbolic space and, finally, in the age of digital technology, as a mapping of our knowledge about the material world in the digital world.

In the context of convergence of digital and physical worlds, ontology is a suitable mechanism to store and convert knowledge. As a universal tool to record knowledge, ontology may do even without formal verbalization of knowledge which makes knowledge digitalization and revisualization simpler. Logical inference based on ontologies allows

* Corresponding author: leshcheva@gsom.pu.ru

digital devices to use recorded knowledge and even generate new knowledge providing a turn of the convergence spiral in the reverse direction.

A detailed practically oriented definition of the term may also be given: "Ontology means a specification or a formal representation of a subject domain which includes an index of terms within the subject domain and logical expressions which describe what the terms mean, how the terms correlate with each other and how the terms may or may not be interrelated" [5].

This study broadens the standard formal definition of ontology by introducing a new component related to data types. The formal model of ontology shall imply an ordered quintuple $O=<C,T,\Re,A,I>$ where C is the plurality of notions (classes, concepts) which form ontology; T is the plurality of data types; $\Re=<\Re o,\Re d>$ is the plurality of relations, wherein $\Re o \subseteq C \times C$ is the plurality of relations between concepts; $\Re d \subseteq C \times T$is the plurality of properties of concepts; A is the plurality of axioms, i.e. statements on concepts which further allow to make logical inference of other statements; and I is the plurality of individuals (class instances).

Ontology which plurality of axioms is not empty ($A \neq \varnothing$) is called heavyweight, and such ontologies are of the most practical interest since they allow to make inference of new knowledge.

If the plurality of individuals is not empty ($I \neq \varnothing$), such ontology is called an ontology-based knowledge base. Thus, ontology may be divided into two parts. The first part includes ontological knowledge itself that is descriptions of concepts, and relationships between concepts, that is rules describing the subject domain. These rules remain unchanged or change very rarely. The examples of such rules are as follows: "Employee (of an enterprise) is a human", "Head of department X is an employee who manages the operations of the department X". The second part describes particular individuals. For example, "Mikhail Smirnov is Head of department X." This part is subject to constant expansions/changes and may include large amounts of data (data may be evaluations, prices, exchange rates etc. depending on the subject domain).

2 Problem of definition

Information infrastructure of a modern enterprise or company is a complex system. Various technology solutions may be used to support management. Such solutions include both large information systems, covering a full range of business processes and small programs used to solve one routine problem. Some examples of industrial systems are given below:

- ERP (Enterprise Resource Planning)
- MES (Manufacturing Execution System)
- LMS (Learning Management Systems)
- HIS (Hospital Information System)
- LIMS (Laboratory Information Management System)
- CAD (Computer-Aided Design)
- EDMS (Electronic Document Management System).

Examples of a program to solve more specific problems are "working time-keeping system" or "electronic schedule". Moreover, an organisation keeps large amounts of disembodied information which is contained in separate documents (for example, in design documentation, reports, tables etc.) or which is not documented at all.

Tasks requiring retrieval and integration of information from various sources regularly arise in operations management to sample data during the process of making various managerial decisions. For example, combining the data from ERP and MES, consistent roll-out from HIS and LIMS etc. These tasks are often implemented manually or with minimum

automation, although a lot of methods and tools have been developed to solve them. In fact, this is rather difficult to combine all knowledge, information and data of a company into a common knowledge base, e.g. due to segregation of documentation access rights, high rates of new information generation etc. However, it is possible and urgent to define a common development method which would allow creating and maintaining an ontology-based knowledge base to solve highly-specific knowledge integration tasks.

This study is devoted to the description of such methodology: creation and population of an ontology with individuals and relationships between individuals based on structured data stored in various heterogeneous sources. The focus is made on developing an open method which would enable to integrated data of various nature.

3 Approaches to creating ontology-based knowledge bases

There are three basic approaches to creating ontologies and ontology-based knowledge bases [6]:

1. Integration of existing ontologies. During the process of integration, an attempt is made to identify common data in ontologies describing the same or similar subject domains to create a new ontology. Several methods were suggested, namely:
   - merging of ontologies for the purpose of creating a common consistent ontology
   - alignment of ontologies by identifying relationships between them which will allow them to repeatedly use information from each other
   - mapping of ontologies by determining correspondence between elements of the ontologies.
2. Development of ontology from scratch or expansion (population and enrichment) of the existing ontology generally based on the information retrieved from the subject-oriented content.
3. Specialisation of a common ontology to adapt it to a specific subject domain.

Ontology population is understood as adding individuals, with their properties, to the ontology, and ontology enrichment implies addition of new relationships and axioms which use such relationships.

3.1 Currently available approaches to populating ontologies

It is necessary to specify the peculiarities of data integration into knowledge bases. Two approaches are used depending on the problem to be solved. The first approach, mapping on-demand, in similar to ontology mapping, but the relationship is established between the ontology and data stored in distributed data sources, rather than the two or more ontologies.

The approach has proven itself to be useful in the context of very large arrays of decentralised data. It ensures that return values are always relevant as no data copying is done into the ontology. Moreover, it provides the observance of access control policies implemented in the database management system. However, when a large number of data sources is integrated or logical inference machines are used to process complex requests, the performance may become extremely low [7]. A supposition was made [8] that expressiveness of requests to a knowledge base should be limited, as some request types may cause exponential growth in the number of generated requests to data sources.

The second approach is data inclusion (consolidation) in the knowledge base as values of properties of the objects (data materialization). It is used to build data warehouses. After consolidation is completed, all necessary data are put into the ontology and this allows to take all the advantages of knowledge bases as compared with databases, namely logical inference tools and means of visualisation of the relationships between objects. However, this approach conflicts with the principle of separation of ontology code from data and this

makes it difficult to develop and, mainly, maintain the ontology. When a new individual appears or properties of the previously described individual are changed, it its necessary to create/find the respective object in the knowledge base and set/modify values of its properties. To solve the problem of data deterioration, the process of consolidation must be regularly started. That is why, it is necessary to find a compromise between expenses on data updating and sensitivity of the application to deteriorated data. The second limitation may be an amount of data because the obtained knowledge base virtually doubles the data volume. There are Direct Mapping and Domain Semantics-Driven Mapping depending on what structure, the structure of ontology or the structure of data sources is primary or more critical [9].

Direct Mapping is implemented automatically, and the resulting ontology completely inherits the structure of the data source. But the structure of the data source may rarely serve a good description of the subject domain [10]. That is why the result of Direct Mapping is used as a starting point for developing the ontology. Otherwise, Domain Semantics-Driven Mapping is applied.

When the second approach is used, the ontology of the subject domain serves as a basis and the rules for ontology population are established using a certain mapping description language processed by a specialised (pre)processor.

3.2 Sources of data to be integrated

Main types of sources of structured data are relational databases, XML documents, electronic tables and text files with a proprietary structure. Methods of integration of data from various source types are currently being developed in parallel. Let us consider main of them.

Translation of relational databases into ontology. Relational databases are mainly used to store structured data sources. For Direct Mapping, the database schema is directly translated into ontology using a special translating program. The rules of mappings are established by developers, so if different programs are used, a wide range of similar but not identical ontologies may be obtained from one database.

When the database schema is transformed into ontology with account of the subject domain semantics, the mapping rules are explicitly set using mapping description languages. Such languages may be divided into two groups. Languages of the first group are mainly rely on SQL queries to describe data mapping and this is a potential disadvantage because such languages cannot be used to describe complicated cases of mapping. However, SQL popularity facilitates the acceptance of this approach because there is no need to learn a new language. Second group languages use specialised mapping description languages, hence, they may be developed/expanded so that any specific needs are met, for example, keyword search or regular expression search. In practice, expressiveness of the second group languages is very limited [9].

Main features to be provided by mapping languages and tools to implement them:

1. Generation of unique identifiers determined by the user (ability to generate Uniform Resource Identifiers (URI) along with a simple use of primary key values, for example, by combining values in columns etc.)
2. Logical tables (ability to read tuples not only from tables but also from the SQL view or SQL query result).
3. Field selection (ability to select only a subset of table columns for translation).
4. Field renaming (ability to map a column in the RDF property with another name).
5. Selection condition (ability to translate a subset of table tuples only by setting the WHERE conditions in the SELECT operator).
6. Use of the existing ontology (ability to map relational objects into instances of

existing ontologies).

7. Mapping of one table to n classes (ability to use values of a column as a categorization template: tuples of the table will be translated into instances of various ontological classes based on the value of this field).
8. Conversion of the many-to-many relationship into simple triples (ability to translate a combined table representing a many-to-many relationship into a simple triple, unlike direct mapping where the combined table will be translated into a separate class).
9. Creation of blank nodes (ability to create blank nodes and refer to them within a graph obtained in the translation process).
10. Translation of data types (ability to process relational data types into RDF data types).
11. Data processing (ability to apply the conversion function to the values prior to formation of RDF triples, for example, complex conversion of a type like computation of a value using several columns, line processing etc.)

The urgency of the issue of populating ontologies with data from relational databases with account of the subject domain semantics has resulted in development of a wide range of mapping description languages, such as DR2 MAP [11], R2O [12], DR2Q [13]. We should also mention languages which were created and developed within one and the same project, for example, METAmorphoses [14], RDBToOnto [15], Relational.OWL [16], but they are no longer supported.

In 2012, the W3C RDB2RDF Working Group issued two Recommendations: A Direct Mapping of Relational Data to RDF [17] and R2RML: RDB to RDF Mapping Language [18]. As it follows from its title, the first Recommendation regulates direct mapping of relational databases to RDF. The other document defines a language for describing mapping from a relational database to RDF but includes no guidance on implementation of the R2RML processor, that is why there is a number of R2RML compatible tools with unique approaches to implementation, for example, Ultrawrap [19], DB2Triples [20]. The R2RML uses the best features of its predecessors: it is developed as a result of learning of its predecessors, is based on their experience and includes most of their expressiveness. That is why; it seems that using of the R2RML is inevitable when it comes to conversion of relational databases into RDF. Moreover, tools appeared earlier than R2RML also support it now. For example, Virtuoso [21] includes a simple adapter which translates R2RML in syntax of its own language. However, the authors [9] believe that the R2RML cannot be considered as a comprehensive solution, so new suggestions will appear. This is due to the fact that some approaches rely on complex templates which cannot be expressed using the R2RML language. For example, RDBToOnto [15] analyses data redundancy to define categorisation templates. The publication [22] describes an intelligent data analysis method used to automatically define mappings between the database and the existing ontology. And, finally, the R2RML does not provide data manipulation functions. It relies on the functionality of a relational DBMS, but not on its internal interface.

Integration of data from XML documents. XML (Extensible Markup Language) [23] is an extensible markup language, a simple flexible Internet-oriented textual format developed to structure, store and transmit information between applications. XML documents may be divided into two categories, such as data-oriented and content-oriented categories, but only the first one is considered in the context of XML schema conversion to ontology.

First methods of converting XML documents to an ontology [24, 25] implemented Directed Mapping, then the methods allowing to convert XML schemas to an ontology, for example [26, 27], were developed. One of the well-developed methods as regards to the functionality provided by them is described in [28]. This method enables, in particular, to map several XML schemas to one existing ontology, including creation of individuals. For

this purpose, the authors have developed their own mapping language which is used to describe how XML nodes are converted to ontology elements.

It should be noted that none of the suggested methods could become a standard and each project where XML document conversion to ontology is required uses its own tools especially developed to meet the needs of the project and to consider its peculiarities.

Conversion of electronic tables into ontology. Despite of the fact that electronic tables are a prevailing tool to represent and process structured data, the issue of converting tables into ontology is relatively rare discussed. We may suggest that it is so due to the fact that tables may be represented in the form of relational databases for which a number of methods have already been developed. However, for projects where many separate tables of different structures are used, an extra conversion may become a significant limitation. That is why; it is practical to develop a method for converting tables into ontology.

The publication [29] suggests TANGO (Table ANalysis for Generating Ontologies) which is explained as follows: firstly, tables are "canonized", i.e. tables are converted into a standard form suitable for further analysis; secondly, a set of "mini-ontologies" is created where each "mini-ontology" describes the structure of one canonized table; and thirdly, the "mini-ontologies" are merged into one resulting ontology. Thus, TANGO implements Directed Mapping, and the ontology created as a result of TANGO application describes the structure of tables and, in general, cannot be considered as ontology of a subject domain.

3.3 Limitations and disadvantages of existing solutions

A vast number of approaches to integrating data into ontologies have been suggested. However, all of them have some limitations. The disadvantages of the existing solutions include:

1. The need to use and, accordingly, to learn the syntax of several specific mapping description languages. In addition, a convenient tool must be created for each of the languages being used to ensure effective performance.
2. A vast number of various methods for integrating data to RDF have bee suggested, but each method is suitable for one data source type only.
3. It is unclear how simultaneously integrate data from multiple sources. If it is necessary to integrate data, for example, from a database or a XML document into one ontology, the operations should be carried out consequently, using existing methods. For example, at first, populate the initial ontology with data from the DB and then populate the ontology obtained at the previous step with the data from the XML document. Due to constant modifications of data in sources, the above mentioned procedure must be regularly repeated and that will cause unreasonable expenses.
4. Conflicts may arise during integration of data from various sources. The authors of the above mentioned method focus on integration of data from one source, so the ways to resolve such conflicts are not discussed.
5. It is unclear how the connection with the data source is established, how authentication is carried out etc.
6. Companies may use proprietary data storage formats and new formats may be created in the future. No recommendations concerning data integration in this case are given.

The problem analysis allows concluding that it is necessary to develop a new method which will enable to integrate data from various source types. Development of such method will reduce the complexity of populating and enriching ontologies using the collected data arrays, irrespective of the format in which data are stored or represented.

4 Suggested method

A method of automated population and enrichment of ontologies or ontology-based knowledge bases by consolidation based on structured data stored in various heterogeneous sources is suggested. Main types of structured data sources are relational databases, XML documents, electronic tables and text files with a proprietary structure. The algorithm of data consolidation process is shown in Fig. 1.

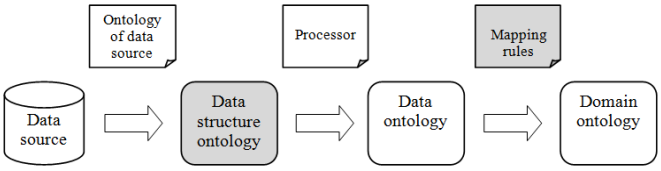

Fig. 1. Algorithm of data consolidation process.

In order to ensure processing of data from various source types, ontologies generally describing the structure of such sources have been created. For example, ontology of a relational database is shown in Fig. 2 in a simplified form. It should be noted that ontologies are a considerably expressive tool to describe structures of data of any type and format.

Using the ontology of data source type, a developer describes the structure of the source which data have to be put in the knowledge base. A special processor program generates an intermediate ontology based on the ontology of data structure; such intermediate ontology including required data from the described source.

The developer applies the mapping rules to describe in detail how the data will be integrated in the ontology. The rules may also govern integration process behaviour if any conflicts arise in data from different sources. Due to application of the rules, ontology of the subject domain is enriched by the data from intermediate ontologies.

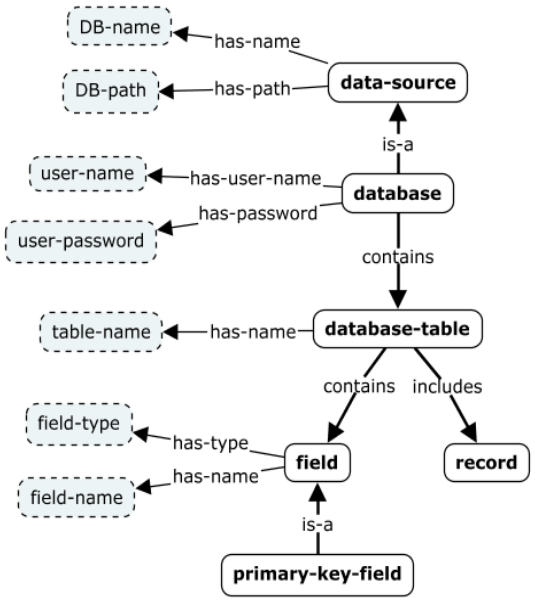

Fig. 2. Ontology of a relational database.

The suggested method has a modular architecture and enables to integrate data from different source types and to consider the distributed nature of data and the necessity of authentication during accessing network resources. The method also suggests a solution to the problem of integrated data conflict resolution so that the complexity of populating and enriching ontologies is reduced using the collected data arrays, irrespective of the format in which data are stored or represented. The method is devoid of the majority of the above mentioned disadvantages. In particular, a developer may describe ontology of the data structure and mapping rules using ontology tools known to him/her. The suggested solution is also extendible to provide a capability to be used with new data formats.

5 Conclusion

The problem of creation of integrated and automated knowledge management systems becomes more and more relevant for companies to ensure that a company's intelligent assets are stored and critical business processes are supported.

Effectiveness of managerial decisions depends, in particular, on completeness and consistency of the information available as well as the possibility to process it flexibly, including at the semantic level. Business processes are often automated using software solutions of in-house and outsourced developers without taking into account of their interrelationships; data are stored in different formats and different sources. Due to constant changes in business processes, knowledge management system developers have to correct data models that results in structural and semantic heterogeneity of information elements. The use of such solutions leads to an increase in complexity and, therefore, reduces effectiveness and convenience of knowledge management systems.

Ontologies and knowledge bases may be used to describe not only business processes but also products, documents, competences, technologies, functions, strategies, financial flows etc. This tool is oriented on interoperability which means the possibility to be used at different hierarchical levels, in different departments, on different hardware and software platforms by differently qualified staff. The aggregate of corporate ontologies serve as a universal framework of the company and, simultaneously, a bridge to understand and transfer knowledges and technologies. But generating such ontologies is a first step only. All advantages of ontology-based knowledge bases are revealed only when they are populated and enriched with data arrays collected by the company.

This study is supported by grant no. 17-07-00228 issued by the Russian Foundation for Basic Research.

References

1. A. L. Dobrokhotov, *New Encyclopaedia of Philosophy* (Mysl, Moscow, 2010)
2. R. Studer, et al., *Knowledge Engineering and Agent* Technology (IOS Press, Amsterdam 2000)
3. T. R. Gruber, Knowledge acquisition**, 5(2)** (1993)
4. M. R. Genesereth, N. J. Nilsson, *Logical Foundations of Artificial Intelligence* (Morgan Kaufmann, Los Altos, 1987)
5. T. A. Gavrilova, D. I. Muromtsev, *Intelligent Technologies in Management – Tools and Systems* (Hidher School of Management, St. Petersburg, 2007)
6. G. Petasis et al., *Knowledge-driven multimedia information extraction and ontology evolution* (Springer-Verlag, Berlin, 2011)

7. S. S. Sahoo et al., W3C RDB2RDF Incubator Group Report [online], Available at: http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport_01082009.pdf (2009)
8. O. Erling, Requirements for relational to RDF mapping [online], Available at: http://www.w3.org/wiki/Rdb2RdfXG/ReqForMappingByOErling (2008)
9. F. Michel, J. Montagnat, C. Faron-Zucker A survey of RDB to RDF translation approaches and tools, Research report [online], Available at: http://hal.archives-ouvertes.fr/hal-00903568 (2014)
10. C. Dolbear, J. Goodwin, *W3C Workshop on RDF Access to Relational Databases*, (2007)
11. C. Bizer, *Proceedings of the 12th International World Wide Web Conference* (2003)
12. J. Barrasa, Ó. Corcho, A. Gómez-pérez, *Proceedings of the 2nd Workshop on Semantic Web and Databases* (2004)
13. R. Cyganiak, C. Bizer, O. Maresch, C. Becker, The D2RQ mapping language v0.*8* [online], Available at: http://d2rq.org/d2rq-language (2012)
14. M. Svihla, I. Jelinek, *Proceedings of Electronic Computers and Informatics (ECI)* (2004)
15. F. Cerbah*, Learning highly structured semantic repositories from relational databases* (Springer Berlin, Heidelberg, 2008)
16. C. P. De Laborda, S. Conrad, *Conceptual Modelling-ER 2006* (Springer Berlin, Heidelberg 2006)
17. A Direct Mapping of Relational Data to RDF [online], Available at: https://www.w3.org/TR/rdb-direct-mapping/ (2012)
18. R2RML: RDB to RDF Mapping Language [online], Available at: https://www.w3.org/TR/r2rml/ (2012)
19. Ultrawrap [online], Available at: http://capsenta.com/ultrawrap (2014)
20. DB2Triple*s* [online], Available at: https://github.com/antidot/db2triples (2018)
21. Virtuoso [online], Available at: http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/ (2016)
22. W. Hu, Y. Qu, Discovering simple mappings between relational database schemas and ontologies, Springer Berlin Heidelberg, 225-238 (2007)
23. M. Ferdinand, C. Zirpins, D. Trastour, *Lifting XML Schema to OWL, Web Engineering* (ICWE 2004, Munich, 2004)
24. R. García, O. Celma, *Semantic Integration and Retrieval of Multimedia Metadata* (ISWC, Galway 2005)
25. N. Anicic, N. Ivezic, Z. Marjanovic, *Mapping XML Schema to OWL, Enterprise Interoperability* (Springer, London, 2007)
26. C. Cruz, C. Nicolle, ODBIS, **2008** (2008)
27. T. Rodrigues, P. Rosa, J. Cardoso, Mapping XML to Exiting OWL ontologies, International Conference WWW (2006)
28. Y. A. Tijerino et al., World Wide Web, **8(3)** (2005)