Poster Paper
Data Integration for Supporting Biomedical Knowledge Graph Creation at Large-Scale

Samaneh Jozashoori[1,2][0000−0003−1702−8707], Tatiana Novikova[3][0000−0003−1334−0998], and Maria-Esther Vidal[2,1][0000−0003−1160−8727]

[1] L3S Institute, Leibniz University of Hannover, Germany
[2] TIB Leibniz Information Centre for Science and Technology, Germany
[3] University of Bonn, Germany
jozashoori@l3s.de
s6tanovi@uni-bonn.de
maria.vidal@tib.eu

Abstract. In recent years, following FAIR and open data principles, the number of available big data including biomedical data has been increased exponentially. In order to extract knowledge, these data should be curated, integrated, and semantically described. Accordingly, several semantic integration techniques have been developed; albeit effective, they may suffer from scalability in terms of different properties of big data. Even scaled-up approaches may be highly costly due to performing tasks of semantification, curation, and integration independently. To overcome these issues, we devise ConMap, a semantic integration approach which exploits knowledge encoded in ontologies to describe mapping rules in a way that performs all these tasks at the same time. The empirical evaluation of ConMap performed on different data sets shows that ConMap can significantly reduce the time required for knowledge graph creation by up to 70% of the time that is consumed following a traditional approach. Accordingly, the experimental results suggest that ConMap can be a semantic data integration solution that embody FAIR principles specifically in terms of interoperability.

1 Introduction

With the rapid advances in different techniques in the biomedical domain such as Next Generation Sequencing [9], which allow for producing massive amounts of data in acceptable time, and access policies such as FAIR [10] and open data principles, big data has become a quotidian occurrence. However, knowledge discovery from big data, as the criteria to make decisions and take actions, is still a challenging problem. In order to extract knowledge, data residing in different sources should be curated, integrated, and semantically described.

In recent years, the development of Semantic Web technologies with the main purpose of describing the meaning of data in a machine readable fashion, has

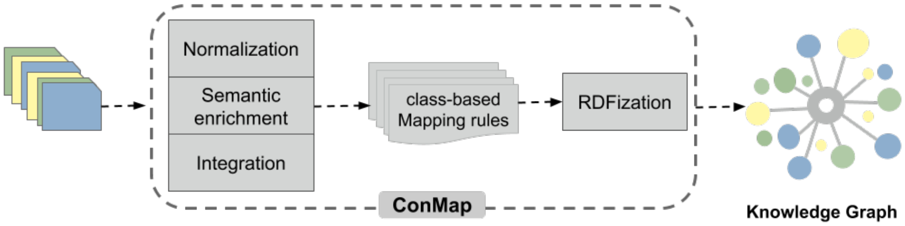


Fig. 1. The ConMap Approach. ConMap receives structured data sets from heterogeneous sources as input, and produces a knowledge graph. It relies on conceptual or class-based mapping approach in performing all tasks of semantic enrichment, integration, and transformation i.e. both semantification and integration are performed during class-based mapping and afterwards, based on generated mapping rules, normalized data is transformed as RDF model into the knowledge graph.

facilitated the implementation of various semantic integration applications. Existing semantic data integration approaches rely on a common framework that allows for the transformation of data in various raw formats into a common data model, e.g., RDF [8]. The mapping rules for data transformation are expressed using mapping languages such as RML [1]. Accordingly, several semantic data integration approaches and tools have been introduced following these technologies such as Karma[4], MINTE [2], SILK [4], and Sieve [7]. Albeit effective, existing semantic data integration tools may suffer from scalability in terms of the dominant dimensions of big data, i.e., volume, variety, veracity, velocity, and value. In fact, even scaled-up approaches are mainly a trade-off between mentioned aspects of big data since it would be highly costly to scale up all the tasks of semantification and integration in terms of more than one dimension, particularly, while being performed independently. More precisely, performing the tasks sequentially results in going through the same procedure of cleaning, semantifying, curating, and transforming for each single data set while their data overlap partially. Moreover, during each mentioned step, the volume of data may be grown and consequently the computational complexity of the next task in the queue, and eventually integration as the last step, will be considerably increased. To overcome drawbacks of existing approaches, we introduce ConMap, a semantic integration approach for big data.

ConMap exploits knowledge encoded in a global schema to perform all the mentioned tasks, i.e., semantification, integration, and transformation in one single step. Therefore, ConMap can be considered as a scalable solution for semantic integration of big data. We have performed an initial experiment study over data sets of various sizes. The observed results suggest that ConMap reduces RDFization time [5] i.e., the time required for transforming heterogeneous structured data sets into RDF.

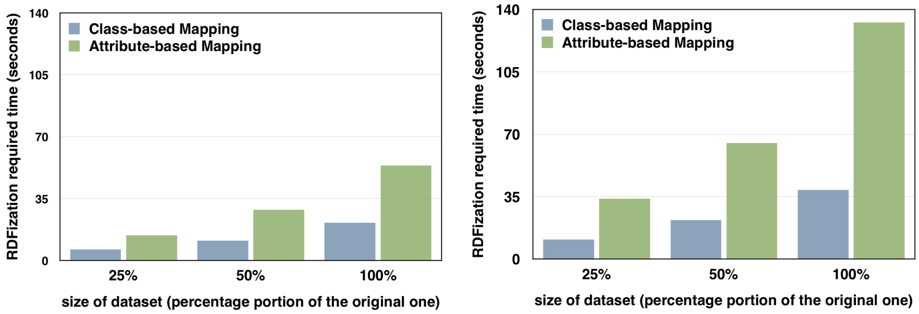
[4] http://usc-isi-i2.github.io/karma/

The rest of the paper is structured as follows: in Section 2 the general idea of ConMap is presented as well as detailed explanation of ConMap architecture and components. In Section 3, the experiments that are performed between ConMap and the attribute-based mapping approach are described and the results are evaluated in terms of time complexity. Finally, Section 4 represents our conclusions.

2 The ConMap Approach

ConMap is a semantic data integration approach able to use mapping rules not only for data semantification, but also for curation and integration. ConMap implements a class-based mapping paradigm that resembles the Global-As-View [3] approach of data integration systems [6]; it enables the definition of the mapping, curation, and integration rules per each class in the global schema. Thus, ConMap executes all the tasks, i.e., semantification, curation, and integration at the same time by evaluating these class-based rules. Figure 1 devises the ConMap architecture. ConMap receives real-world data source(s) that represent the same concepts in the global schema but in different formats; it outputs a knowledge graph where input data is integrated and described in a structured way. Data related to each class is extracted from different data sources which are normalized in advance to reduce data redundancies. Afterwards, normalized data is semantified in order to describe and integrate this data in the knowledge graph. The components of ConMap can be summarized as below:

- **Normalization:** To overcome interoperability issues, all data sets are normalized considering the real world concepts. Since each concept may be represented by more than one data source and each data source may involve more than one concept, the process of normalization is based on the decomposition of each data set in terms of the global schema classes.
- **Semantic Enrichment:** Since the mapping process enables semantification and curation of data, the approach that is applied for mapping plays a significant role both on computational complexity and the quality of semantified data. The attribute-based mapping approach is source-oriented which means it semantifies the concepts that are presented in each source based on the attributes that are available in the same data source. In contrast, the class-based mapping approach is concept-oriented; it defines semantic descriptions of each concept according to the attributes that are determined by a global schema and expressed by variant data sources either equally or differentially.
- **Integration:** To integrate data residing at heterogeneous sources, they all required to be transformed into a unified representation. Since in ConMap approach, in order to decrease the cost of data comparison, data integration precedes data transformation, semantic descriptions provided during the generation of mapping rules are employed to translate different representations of data into a unified one. Furthermore, semantic descriptions provide actionable information for data fusion in case of inconsistency of data values between different sources.



(a) RDFizing one class including 5 attributes from three different sized data sets

(b) RDFizing one class including 12 attributes from three different sized data sets

Fig. 2. Experimental results. (a) The required time for RDFization of one class including five attributes from three different sized data sets. (b) The required time for RDFization of one class including twelve attributes from three different sized data sets.

- **RDFization:** The last component to be executed in ConMap is RDFization. It is important to note that despite being the last step, the performance of previous steps can be also observed through the outcome of this component which evaluates class-based mapping rules for transforming normalized and semantified data sets into the knowledge graph.

3 Experimental Study

In this paper, the performance of two mapping paradigms are compared: the class-based mapping approach provided by ConMap, and an attribute-based approach which is commonly followed by existing tools, e.g., Karma. We address two research questions: **RQ1)** Does ConMap reduce the time complexity of RDFization? **RQ2)** How influential a mapping approach can be in terms of execution time when the complexity of the class increases?

Benchmark: In this study, a data set with overall size of 169.8 MB is extracted from COSMIC[5], an online database of somatic mutations that are found in human cancer. The data set is in tab separated format comprising 557,162 records of lung cancer related coding point mutations that are derived from targeted and genome wide screens.

Metrics: The behavior of the studied mapping approaches is evaluated by measuring the execution time in seconds for transforming a data set into RDF applying that approach.

Implementation: The mapping rules[6] are expressed in the RML mapping language. The RDFization is implemented in Python 3.6. The experiment was

[5] https://cancer.sanger.ac.uk/cosmic

[6] https://github.com/samiscoding/DILS

executed on a Ubuntu 17.10 (64 bits) machine with Intel W-2133, CPU 3.6GHz, 1 physical processor; 6 cores; 12 threads, and 64 GB RAM.

Experimental Setup: Two experiments are set up in this study: **E1)** In order to better understand the influence of mapping approach on time complexity of RDFization, the experiment is run on three different sized data sets: the first one is the preprocessed data set derived from the original mutation data set that is extracted from COSMIC without any decrease regarding its size while the two other data sets are extracted from the first one. The records that are included in two latest data sets are 50% and 25% randomly selected records of the first data set. The result of this experiment is shown in Figure 2(a). **E2)** To study how time complexity of each mapping approach fluctuated with the increase in the number of attributes for a class, for each mapping approach two separated sets of mapping rules are defined; one mapping rule set for an RDF class with twelve attributes and the other one including five attributes. The experimental results can be seen in Figure 2(b). Based on the results of explained experiments that are illustrated in Figure 2, the execution time increases in case of using the attribute-based mapping rules for transformation of data in both sets including different numbers of attributes which positively answers the **RQ1** . Moreover, the observed results lead to answer **RQ2** as follows: in attribute-based mapping approach, the required execution time for transforming one class of data will grow when the number of its attributes increases, however, in class-based mapping the time complexity is not a function of class complexity.

3.1 Discussion

The evaluation results can be simply explained according to the fact that the attribute-based mapping approach performs the same procedure of creating *subject-predicate-object* triple for every single attribute of a class. In contrast, the class-based mapping approach transforms each concept or class including all its attributes to one RDF class in a single run. Therefore, class-based mapping approach can be considered as a fundamental procedure for transforming raw data into RDF model in an integrated non-redundant way.

4 Conclusions an Future Work

We introduced ConMap, a semantic integration approach that deploys the knowledge encoded in an ontology to perform semantification and integration during transformation, in a way that a big data scalability can be acquired. We empirically showed that ConMap can reduce the execution time of semantic integration of structured data sets into a knowledge graph. Although experimental results demonstrated in this paper were derived by all components of ConMap, there is still room to illustrate the power of this approach in terms of integration. There exist open problems regarding different dimensions of big data that can be tackled during the integration process in ConMap; they include veracity which refers to noise, abnormality and inconsistency of available data. In our future work we will more focus on improving ConMap from data fusion perspective.

5 Acknowledgement

This work has been supported by the European Union's Horizon 2020 Research and Innovation Program for the project iASiS with grant agreement No 727658.

References

1. A. Dimou, M. Vander Sande, P. Colpaert, R. Verborgh, E. Mannens, and R. Van de Walle. Rml: A generic language for integrated rdf mappings of heterogeneous data. In *LDOW*, 2014.
2. K. M. Endris, M. Galkin, I. Lytra, M. N. Mami, M.-E. Vidal, and S. Auer. Mulder: querying the linked data web by bridging rdf molecule templates. In *International Conference on Database and Expert Systems Applications*, pages 3–18. Springer, 2017.
3. M. Friedman, A. Y. Levy, T. D. Millstein, et al. Navigational plans for data integration. *AAAI/IAAI*, 1999:67–73, 1999.
4. R. Isele and C. Bizer. Active learning of expressive linkage rules using genetic programming. *Web Semantics: Science, Services and Agents on the World Wide Web*, 23:2–15, 2013.
5. A. Jha, Y. Khan, M. Mehdi, M. R. Karim, Q. Mehmood, A. Zappa, D. Rebholz-Schuhmann, and R. Sahay. Towards precision medicine: discovering novel gynecological cancer biomarkers and pathways using linked data. *Journal of biomedical semantics*, 8(1):40, 2017.
6. M. Lenzerini. Data integration: A theoretical perspective. In *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 233–246. ACM, 2002.
7. P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.
8. E. Miller. An introduction to the resource description framework. *Bulletin of the American Society for Information Science and Technology*, 25(1):15–19, 1998.
9. J. S. Reis-Filho. Next-generation sequencing. *Breast Cancer Research*, 11(3):S12, 2009.
10. M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3, 2016.