Semantic Technologies in the Public Administration: Data and Process Management at unibz

Diego Calvanese, Francesco Corcoglioniti, Julien Corman, Davide Lanti, Marco Montali, Alessandro Mosca, Nicolas Troquard, Guohui Xiao
Faculty of Computer Science, Free University of Bozen-Bolzano, Bolzano, Italy
firstname.lastname@unibz.it

Abstract

We report on the major research and project activities carried out at the Faculty of Computer Science of the Free University of Bozen-Bolzano in AI that are relevant for the Public Administration.

Keywords: semantic technologies; ontology-based data access; OBDA; ontology-based data integration; OBDI; virtual knowledge graphs; query rewriting; mapping patterns; process mining; event logs; event log extraction.

1 Introduction

Public administrations are often large organizations that have to deal with large and complex collections of data. Such data might come from different data silos inside the organization, e.g., from different branches, entities or offices inside the administration, but also through the exchange of data with other administrations. To be effectively used, accessed, and queried in a uniform way, the data need to be integrated, which requires on the one hand to overcome heterogeneity at various levels, and on the other hand to clean, de-duplicate, and homogenize the data. These are work-intensive, expensive, but essential activities. They might also need to be complemented by enriching the data with additional information, e.g., coming from external sources. Moreover, understanding and managing the complex processes underlying the activities carried out in public administrations, is a complex task, especially when considering how these processes affect and are affected by the underlying data.

We report here on the major research and project activities in the area of semantic technologies in Artificial Intelligence that are carried out at the Faculty of Computer Science of the Free University of Bozen-Bolzano (unibz) and that are relevant for the Public Administration.

2 Involved Researchers

At the Faculty of Computer Science of unibz, research on semantic technologies for data and process management is carried out within two of the four faculty research centres, specifically at the *Research Centre for Knowledge and Data* (KRDB), and at the *Smart Data Factory* (SDF). The latter is actually a technology transfer lab for the Faculty, located at the NOI Technology Park in Bolzano.

The people involved in the reported research activities are: Diego Calvanese (KRDB, SDF), Francesco Corcoglioniti (KRDB), Julien Corman (KRDB), Davide Lanti (KRDB), Marco Montali (KRDB), Alessandro Mosca (SDF), Nicolas Troquard (KRDB), and. Guohui Xiao (KRDB).

3 Research Themes

The research in semantic technologies carried out at unibz covers two broad areas, respectively dealing with data management and with process management.

3.1 Semantic Technologies for Data Management

The challenges related to data management have been tackled by relying on ontologies, more precisely on an approach that traditionally has been called *ontology-based data management* (OBDM), covering in particular data access (OBDA) and integration (OBDI), see, e.g., the recent survey by Xiao *et al.* [2018a], and that more recently has been rebranded as *Virtual Knowledge Graphs* (VKGs) [Xiao *et al.*, 2019].

VKGs is a paradigm for data access and integration that allows one to overcome the difficulties of traditional approaches based on the relational model, by combining three key ideas, which are also reflected in its name: *(i) data virtualization* (V), which is achieved by avoiding to expose end-users to the actual data sources, and by presenting them instead a conceptual representation of the domain of interest, which is kept virtual and defined by means of suitable *mappings over the sources*; *(ii)* the data are structured in the form of a *graph* (G), which provides more flexibility compared to traditional relational tables; in such graph, domain objects and data values are represented as nodes, and properties of objects are encoded as edges; *(iii)* the graph is enriched by *domain knowledge* (K) represented through an *ontology*, capturing, e.g., concept and property hierarchies, domain and range of properties, and mandatory properties, and allowing one to derive new knowledge from the explicitly asserted one.

Unibz has contributed to laying the theoretical foundations for ODBA/VKGs in a series of seminal works [Calvanese *et al.*, 2007; Poggi *et al.*, 2008; Artale *et al.*, 2009; Calvanese *et al.*, 2006; Calvanese *et al.*, 2013], which was initiated with an article by Calvanese *et al.* [2005] at AAAI 2005, which in 2021 received the *Classic Paper Award* at the 35th AAAI Conf. on Artificial Intelligence. This work led to the development of the *DL-Lite* family

of lightweight ontology languages [Calvanese *et al.*, 2007; Artale *et al.*, 2009], specifically optimized for efficient query answering over (relational) data sources, and later standardized by the W3C as the OWL 2 QL profile of the OWL 2 ontology language. It also led to the proposal of a framework for accessing data through a *DL-Lite* ontology via *declarative mappings* to relational data sources [Poggi *et al.*, 2008].

The foundational research has been complemented by more applied research, aimed at the development of a system implementing efficient query answering over *OWL 2 QL* ontologies via mappings. Specifically, unibz has been working over the past 10 years on the VKG system Ontop, which is a state-of-the-art query reformulation engine for VKGs [Calvanese *et al.*, 2017a; Xiao *et al.*, 2020]. Ontop allows one to efficiently process a query posed over a (domain) ontology expressed in OWL 2 QL, by rewriting the query with respect to the ontology axioms, and unfolding the rewritten query with respect to the data source mappings. A key challenge in VKGs and in Ontop has been to make such query reformulation process highly efficient, generating a query that not only is small but that can also be executed efficiently by the underlying data source. For this purpose, Ontop relies on advanced query optimization techniques, which make extensive usage both of domain knowledge encoded in the ontology, and of constraints on the data sources to which the ontology is mapped [Rodriguez-Muro and Calvanese, 2012; Lanti *et al.*, 2017; Xiao *et al.*, 2018b].

Driven by the need of evaluating the performance of the novel query optimization techniques studied for Ontop, and the lack of proper evaluation and benchmarking frameworks, *benchmarks for the VKG/OBDA setting* have been developed, which properly take into account domain knowledge encoded in the ontology [Lanti *et al.*, 2015; Lanti *et al.*, 2019].

3.2 Semantic Technologies for Process Management

For contemporary organizations, and in particular public administrations, it is increasingly important to analyze how their business processes are executed in the real world, towards quality assurance, optimization, and continuous improvement. An effective framework to tackle this need is provided by *process mining*, where insights are automatically extracted from event data that represent the footprint of process executions inside the organization, and used to discover and enrich process models, provide operational support, check compliance, analyse bottlenecks, and suggest improvements. The applicability of process mining depends on the one hand on the availability of high-quality event data, i.e., process logs recording which process instances have been executed and which events occurred when, and on the other hand on the representation of such data in a format that is understandable by process mining algorithms.

Research at unibz has tackled this issue by leveraging the VKG/OBDA approach to partially automatize the extraction of event logs from legacy information systems. An approach, called *onprom*, has been devised where humans only provide some key conceptual information related to the extraction, based on which the log extraction process is then handled in a fully automatic way. Specifically, the provided conceptual information concerns the relevant concepts and relations, how these map to the underlying information system, and which concepts/relations relate to the key notions of process case, event, and event attributes, as defined in the IEEE eXtensible Event Stream (XES) standard. The conceptual information is provided by suitably (graphically) annotating, with the key elements defined by XES, a conceptual domain model, which in turn is mapped through VKG/OBDA mappings to the underlying legacy data sources [Calvanese *et al.*, 2017d; Calvanese *et al.*, 2017b]. The approach has also been implemented in a tool-chain consisting of three components: *(1)* a UML Editor, used to design the domain ontology; *(2)* an Annotation Editor, allowing domain expert to specify the event data annotations over the domain ontology; and *(3)* a Log Extractor, used to extract from the underlying database the XES event log. Each component can be used both as a plug-in for the extensible process mining framework ProM, or within an integrated toolkit. The produced logs are fully compliant with the XES standard [Calvanese *et al.*, 2017c].

In the approach proposed in the onprom framework, the conceptual elements of XES, which determine the possible types of annotations over the conceptual domain model, are essentially "hardcoded". Such an approach has then been generalized, by observing that the XES ontology can be substituted with an arbitrary (possibly upper level) ontology, which in turn dynamically determines the possible annotation types [Calvanese *et al.*, 2018].

4 Relevant Projects

KAOS: Knowledge-Aware Operational Support:
Duration: 01/06/2016–30/05/2019 (36 months);
Budget: € 110,250 for unibz, € 327,862 total;
Funding: European Region Tyrol-South Tyrol-Trentino (EGTC), Interregional Project Networks (IPN);
Goal: To devise a new generation of operational support techniques empowered with domain knowledge, truly flexible and able to assist humans in the effective execution of business processes inside an organization domain.

IDEE: Data Integration for Energy Efficiency:
Duration: 01/10/2018–30/05/2022 (42 months);
Budget: € 226,393 for unibz, € 394,060 total;
Funding: ERDF 2014-2020.
Goal: To develop a technological infrastructure based on semantic technologies for the integration of data, with an emphasis on energy-related data for buildings, and to provide techniques and tools for data visualization and analysis.

HOPE: High quality Open data Publishing and Enrichment:
Duration: 29/08/2019–28/02/2023 (42 months);
Budget: € 155,332 for unibz, € 808,030 total;
Funding: MIUR, PRIN – Bando 2017;
Goal: To overcome the main technical problems that current open data solutions suffer from, by developing a methodology and tools for a new way of producing, publishing, maintaining, accessing and exploiting privacy-preserving open data.

HIVE: Heterogeneous Data Integration into Virtual Knowledge Graphs:
Duration: 15/04/2021–14/04/2022 (12 months);

Budget: €52,400;
Funding: "Fusion Grant" project sponsored by Fondazione Cassa di Risparmio di Bolzano and Ontopic s.r.l. In coordination with NOI Techpark;
Goal: To overcome the limitations of current solutions for VKGs, which focus on relational databases and are unsuited to applications with heterogeneous data, and extend VKGs to two widely used non-relational data sources, namely web APIs and (structured data from) text documents.

5 Resources and Applications

Ontop is a state-of-the-art Virtual Knowledge Graph system. Ontop translates SPARQL queries expressed over an OWL 2 QL ontology into SQL queries executed by the relational data sources. Ontop is compliant with all relevant Semantic Web standards, and supports all major commercial and free relational data sources, including data federation systems (such as Dremio, Denodo, Teiid). Ontop is an open source community effort, co-developed by unibz, and released under an Apache 2 licence. See https://ontop-vkg.org/.

South Tyrol Open Data Hub Knowledge Graph: In a collaboration with NOI Techpark Bolzano, the unibz spinoff Ontopic s.r.l. (see https://ontopic.ai/) is developing a SPARQL endpoint that provides access to the data of the South Tyrol Open Data Hub (see https://opendatahub.bz.it), presented as a knowledge graph, and allows such data to be queried using SPARQL. The Open Data Hub project envisions the development and set up of a portal whose primary purpose is to offer a single access point to all (Open) Data from the region of South Tyrol that are relevant for the economy sector and its actors. The Open Data Hub Knowledge Graph is powered by Ontop, and is available at https://sparql.opendatahub.bz.it/.

IDEE Domain Ontology for Energy Consumption: In partnership with the municipality of Merano (and an energy provider), the IDEE project is developing a domain ontology of energy consumption, and a VKG (powered by Ontop) of energy consumption data in South Tyrol. The VKG portal https://idee.projects.unibz.it/ (restricted access), will be queried by visualization tools to assist municipalities in decisions about energy policies, especially those signatories of the Covenant of Mayors.

OnProm is a tool suite developed within the project KAOS to support the various phases of the OBDA-based event log extraction framework. It consists of various plug-ins of the ProM extensible process mining framework, and it relies on Ontop for the OBDA functionalities. Onprom is available at http://onprom.inf.unibz.it/.

6 Challenges and Perspectives

In our research on Semantic Technologies carried out at unibz, we are exploring various directions that are highly relevant to address the challenging requirements posed by data and process management in the public administration.

Support for Different Types of Data Sources. The OBDA/VKG framework has been investigated mostly for the case of relational data sources. However, we are highly interested in extending it to other kinds of data sources, such as XML and JSON-document databases, key-value stores, graph databases, semi-structured data, and web APIs. In initial work in this direction we have proposed an architecture for a generalized OBDA systems, and have developed an extension of Ontop to access MongoDB, a popular JSON-document database [Botoeva *et al.*, 2016; Botoeva *et al.*, 2018; Botoeva *et al.*, 2019]. Further research is ongoing within the HIVE project to extend the VKG framework to seamlessly access web APIs and semi-structured data stored in annotated text documents. Moreover, we are studying the problem of how to federate multiple, possibly heterogeneous data sources, without necessarily relying on an intermediate layer provided by a data federation engine.

Support for Different Types of Data. Independently on the actual form of the data sources, it is also of great interest to support specific forms of data with a predefined semantics, notably geospatial data and temporal data. We have already extended the Ontop system to support GeoSPARQL, the extensions of SPARQL with geospatial functions, which are translated to PostGIS, see e.g., [Ding *et al.*, 2021] for a notable use case. However, how to efficiently deal with raster data remains a challenging problem that we are currently investigating. Similarly, the temporal dimension of data is important in many application domains, and requires special treatment, e.g., due to the need to coalesce adjacent temporal intervals, or to compute temporal aggregations. We have made an initial proposal for a framework to deal with temporal data [Güzel Kalayci *et al.*, 2019], but many challenging problems remain to be investigated.

Ontology and Mapping Design. A further key challenge in the OBDA/VKG setting is that of designing the ontology and notably the mappings for complex scenarios. This activity is currently carried out mostly manually, although a (semi-)automatic approach would be highly desirable. A promising direction that we are currently exploring is to rely on *mapping patterns* [Calvanese *et al.*, 2020], and to use such patterns to instrument an automatic mapping extraction process [Calvanese *et al.*, 2021a]. The extracted mappings can then be further refined and adjusted by domain experts.

Process Mining with Common Sense. An interesting direction that we are currently exploring is due to the fact that, with the growth of process mining in breadth (variety of covered tasks) and depth (sophistication of the considered process models), event logs need to be augmented by commonsense knowledge to provide a better input for process mining algorithms. This is crucial to infer key facts that are not explicitly recorded in the logs, but are necessary in a variety of tasks, such as understanding the event data, assessing their compliance and quality, identifying outliers and clusters, computing statistics, and discovering decisions, ultimately empowering process mining as a whole [Calvanese *et al.*, 2021b].

References

[Artale *et al.*, 2009] A. Artale, D. Calvanese, R. Kontchakov, and M. Zakharyaschev. The *DL-Lite* family and relations. *JAIR*, 36:1–69, 2009.

[Botoeva *et al.*, 2016] E. Botoeva, D. Calvanese, B. Cogrel, M. Rezk, and G. Xiao. OBDA over non-relational databases. In *Proc. AMW*, volume 1644 of *CEUR,* ceur-ws.org. CEUR-WS.org, 2016.

[Botoeva *et al.*, 2018] E. Botoeva, D. Calvanese, B. Cogrel, and G. Xiao. Expressivity and complexity of MongoDB queries. In *Proc. ICDT*, volume 98 of *LIPIcs*, pages 9:1–9:22, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[Botoeva *et al.*, 2019] E. Botoeva, D. Calvanese, B. Cogrel, J. Corman, and G. Xiao. Ontology-based data access – Beyond relational sources. *Int. Art.*, 13(1):21–36, 2019.

[Calvanese *et al.*, 2005] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. *DL-Lite*: Tractable description logics for ontologies. In *Proc. AAAI*, pages 602–607, 2005.

[Calvanese *et al.*, 2006] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. In *Proc. KR*, pages 260–270, 2006.

[Calvanese *et al.*, 2007] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Tractable reasoning and efficient query answering in description logics: The *DL-Lite* family. *JAR*, 39(3):385–429, 2007.

[Calvanese *et al.*, 2013] D. Calvanese, G. De Giacomo, D. Lembo, M. Lenzerini, and R. Rosati. Data complexity of query answering in description logics. *AIJ*, 195:335–360, 2013.

[Calvanese *et al.*, 2017a] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, and G. Xiao. Ontop: Answering SPARQL queries over relational databases. *Semantic Web J.*, 8(3):471–487, 2017.

[Calvanese *et al.*, 2017b] D. Calvanese, T. E. Kalayci, M. Montali, and A. Santoso. OBDA for log extraction in process mining. In *RW Tutorial Lectures*, volume 10370 of *LNCS*, pages 292–345. Springer, 2017.

[Calvanese *et al.*, 2017c] D. Calvanese, T. E. Kalayci, M. Montali, and A. Santoso. The onprom toolchain for extracting business process logs using ontology-based data access. In *Proc. of BPM-D&DA*, volume 1920 of *CEUR,* ceur-ws.org. CEUR-WS.org, 2017.

[Calvanese *et al.*, 2017d] D. Calvanese, T. E. Kalayci, M. Montali, and Stefano Tinella. Ontology-based data access for extracting event logs from legacy data: The onprom tool and methodology. In *Proc. BIS*, volume 288 of *LNBIP*, pages 220–236. Springer, 2017.

[Calvanese *et al.*, 2018] D. Calvanese, T. E. Kalayci, M. Montali, A. Santoso, and W. van der Aalst. Conceptual schema transformation in ontology-based data access. In *Proc. EKAW*, volume 11313 of *LNCS*, pages 50–67. Springer, 2018.

[Calvanese *et al.*, 2020] D. Calvanese, A. Gal, D. Lanti, M. Montali, A. Mosca, and R. Shraga. Mapping patterns for virtual knowledge graphs (A report on ongoing research). In *Proc. DL*, volume 2663 of *CEUR,* ceur-ws.org. CEUR-WS.org, 2020.

[Calvanese *et al.*, 2021a] D. Calvanese, A. Gal, N. Haba, D. Lanti, M. Montali, A. Mosca, and R. Shraga. ADAMAP: Automatic alignment of relational data sources using mapping patterns. In *Proc. CAiSE*, volume 12751 of *LNCS*, pages 193–209. Springer, 2021.

[Calvanese *et al.*, 2021b] D. Calvanese, S. Lukumbuzya, M. Montali, and M. Simkus. Process mining with common sense. In *Proc. BPM PROBLEMS*, volume 2938 of *CEUR,* ceur-ws.org, pages 45–50. CEUR-WS.org, 2021.

[Ding *et al.*, 2021] L. Ding, G. Xiao, A. Pano, C. Stadler, and D. Calvanese. Towards the next generation of the LinkedGeoData project using virtual knowledge graphs. *J. Web Semantics*, 71:100662, 2021.

[Güzel Kalayci *et al.*, 2019] E. Güzel Kalayci, S. Brandt, D. Calvanese, V. Ryzhikov, G. Xiao, and M. Zakharyaschev. Ontology-based access to temporal data with Ontop: A framework proposal. *AMCS*, 29(1):17–30, 2019.

[Lanti *et al.*, 2015] D. Lanti, M. Rezk, G. Xiao, and D. Calvanese. The NPD benchmark: Reality check for OBDA systems. In *Proc. EDBT*, pages 617–628. OpenProceedings.org, 2015.

[Lanti *et al.*, 2017] D. Lanti, G. Xiao, and D. Calvanese. Cost-driven ontology-based data access. In *Proc. ISWC*, volume 10587 of *LNCS*, pages 452–470. Springer, 2017.

[Lanti *et al.*, 2019] D. Lanti, G. Xiao, and D. Calvanese. VIG: Data scaling for OBDA benchmarks. *Semantic Web J.*, 10(2):413–433, 2019.

[Poggi *et al.*, 2008] A. Poggi, D. Lembo, D. Calvanese, G. De Giacomo, M. Lenzerini, and R. Rosati. Linking data to ontologies. *J. on Data Semantics*, 10:133–173, 2008.

[Rodriguez-Muro and Calvanese, 2012] M. Rodriguez-Muro and D. Calvanese. High performance query answering over *DL-Lite* ontologies. In *Proc. KR*, pages 308–318, 2012.

[Xiao *et al.*, 2018a] G. Xiao, D. Calvanese, R. Kontchakov, D. Lembo, A. Poggi, R. Rosati, and M. Zakharyaschev. Ontology-based data access: A survey. In *Proc. IJCAI*, pages 5511–5519. IJCAI Org., 2018.

[Xiao *et al.*, 2018b] G. Xiao, R. Kontchakov, B. Cogrel, D. Calvanese, and E. Botoeva. Efficient handling of SPARQL optional for OBDA. In *Proc. ISWC*, LNCS, pages 354–373. Springer, 2018.

[Xiao *et al.*, 2019] G. Xiao, L. Ding, B. Cogrel, and D. Calvanese. Virtual Knowledge Graphs: An overview of systems and use cases. *Data Intelligence*, 1(3):201–223, 2019.

[Xiao *et al.*, 2020] G. Xiao, D. Lanti, R. Kontchakov, S. Komla-Ebri, E. Güzel-Kalayci, L. Ding, J. Corman, B. Cogrel, D. Calvanese, and E. Botoeva. The virtual knowledge graph system Ontop. In *Proc. of ISWC*, volume 12507 of *LNCS*, pages 259–277. Springer, 2020.