Institution: Birkbeck, University of London

Unit of Assessment: 11: Computer Science and Informatics

Title of case study: Virtual Knowledge Graphs in industry and the public sector

Period when the underpinning research was undertaken: 2007 to present

Period when the claimed impact occurred: 2012 to present

Is this case study continued from a case study submitted in 2014? No

1. Summary of the impact

Modern organisations accumulate vast amounts of data, stored in multiple and complex databases. Extracting data is a time-consuming and onerous process, especially for non-IT specialists. Virtual Knowledge Graphs (VKGs) provide users with a search vocabulary that facilitates information extraction without relying on IT specialists, leading to cost/efficiency savings and opening up data repositories to data analytics.

Kontchakov and Zakharyaschev's research underpins (1) the W3C standard VKG language **OWL 2 QL**, making it integral to this flourishing research area, which has created dedicated spinoff companies developing VKG systems and benefited numerous commercial/industrial organisations; (2) reasoning algorithms in the VKG system **Ontop**, which has applications across a wide range of sectors including energy, healthcare, education and innovation.

2. Underpinning research

Research in the area of **Virtual Knowledge Graphs** (VKGs) — also widely known as Ontology-Based Data Access — was initiated in the Department of Computer Science and Information Systems in 2007. In a nutshell, VKGs are built over an organisation's existing data repositories to provide users with access to a knowledge graph (KG) based on concepts and relations in the application domain. The KG contains both the data from the repository and the domain background knowledge, which is represented by an **ontology**; database tables and other entities in data repositories are connected to terms in the ontology via a **mapping**. The VKG system translates (rewrites) queries that are expressed using the KG vocabulary into queries over the underlying data repositories, executes the rewritten queries, and returns the results, after first translating them back into the KG vocabulary. Virtualisation of the KG allows data to be stored and maintained in the original repositories. However, the query rewriting process, which takes account of the ontology and mapping, requires extensive query optimisation. Based on standardised languages, VKG systems provide uniform access to relational and NoSQL databases and can be integrated with visual query interfaces and business intelligence and analytic tools. The VKG technology can make vast amounts of complex heterogeneous data easily accessible to end-users not familiar with database management systems and the structure of the data repositories.

Kontchakov and Zakharyaschev began their work on VKGs with a comprehensive investigation of description logics underlying the W3C Web Ontology Language OWL, aiming to pinpoint the exact boundaries that guarantee **rewritability** of all conjunctive queries mediated by ontologies into first-order logic (FO), that is, into SQL queries for relational databases. The results of that research, conducted in collaboration with Artale and Calvanese from the Free University of Bozen-Bolzano, Italy (UniBZ), were published in a seminal article [1] (with 600+ Google Scholar citations) that was used in 2012 by the W3C to define and standardise **OWL 2 QL**, the profile of OWL 2 designed specifically for VKG systems.

In 2012, Kontchakov, Zakharyaschev and Birkbeck postdoctoral researcher Kikot devised and implemented a highly efficient **query rewriting algorithm** [2] and were invited to join the UniBZ team (led by Rodríguez-Muro) developing one of the major VKG systems, Ontop (ontop-vkg.org). The query rewriting algorithm [2] combined with techniques for compiling ontologies into mappings (so-called saturated mappings) and semantic query optimisation (based on database integrity constraints) dramatically improved Ontop's performance [3]. The team, now including Rezk and Xiao (UniBZ), extended these core components of Ontop with non-monotonic features of the standard KG query language SPARQL 1.1 [4].

In a major JACM article [5], in collaboration with Bienvenu (University of Montpellier, France) and Podolskii (Steklov Mathematical Institute, Russia), the Birkbeck team undertook a comprehensive study of the succinctness of rewritings and the computational complexity of answering queries mediated by OWL 2 QL ontologies. For this purpose, they developed a new computational model for Boolean functions, called **hypergraph programs**, which connects

rewritings to circuit complexity, leading to a complete classification of ontology-mediated queries by the complexity of query answering and the size of the query rewritings. In practical terms, this work provides users of VKG systems with guidance and recommendations on the efficiency of their queries.

Kontchakov and Zakharyaschev collaborated during 2016-17 with Hovland, Skjæveland and Waaler (University of Oslo, Norway) to design an ontology and mapping for the **Slegge database** at Equinor (formerly Statoil), which was used in the EU FP7 Integrated Project **Optique**. This resulted in the open-access publication [6] of a complete set of resources (including geologists' information needs and corresponding SPARQL queries) that provides developers of VKG systems with a real-world VKG application that can be utilised as a benchmark for the effectiveness and efficiency of their systems.

3. References to the research

1. A. Artale, D. Calvanese, R. Kontchakov, M. Zakharyaschev. The DL-Lite Family and Relations. *J. Artif. Intell. Res.* (JAIR) 36:1-69 (2009). doi.org/10.1613/jair.2820
2. S. Kikot, R. Kontchakov, M. Zakharyaschev. Conjunctive Query Answering with OWL 2 QL. In *Proc. of the 13th Int. Conf. on Knowledge Representation and Reasoning*, *KR 2012,* pp. 275-285. AAAI Press, 2012. www.aaai.org/ocs/index.php/KR/KR12/paper/view/4538
3. M. Rodríguez-Muro, R. Kontchakov, M. Zakharyaschev. Ontology-Based Data Access: Ontop of Databases. In *Proc. of the 12th Int. Semantic Web Conf., ISWC 2013*, vol. 8218 of LNCS, pp. 558-573. Springer, 2013. doi.org/10.1007/978-3-642-41335-3_35
4. R. Kontchakov, M. Rezk, M. Rodríguez-Muro, G. Xiao, M. Zakharyaschev. Answering SPARQL Queries over Databases under OWL 2 QL Entailment Regime. In *Proc. of the 13th Int. Semantic Web Conf., ISWC 2014*, vol. 8796 of LNCS, pp. 552-567. Springer, 2014. doi.org/10.1007/978-3-319-11964-9_35
5. M. Bienvenu, S. Kikot, R. Kontchakov, V. Podolskii, M. Zakharyaschev. Ontology-Mediated Queries: Combined Complexity and Succinctness of Rewritings via Circuit Complexity. *Journal of the ACM* 65(5) Article 28 (2018). doi.org/10.1145/3191832
6. D. Hovland, R. Kontchakov, M. Skjæveland, A. Waaler, M. Zakharyaschev. Ontology-Based Data Access to Slegge. In *Proc. of the 16th Int. Semantic Web Conf., ISWC 2017*, vol. 10588 of LNCS, pp. 120-129. Springer, 2017. doi.org/10.1007/978-3-319-68204-4_12

Grants:
'quantMD: Ontology-Based Management for Many-Dimensional Quantitative Data' (EPSRC, 2019-22, £453,015)
'iTract: Islands of Tractability in Ontology-Based Data Access' (EPSRC, 2015-18, £354,117)
'ExODA: Integrating description logics and database technologies for expressive ontology-based data access' (EPSRC, 2010-13, £271,714)

4. Details of the impact

In 2012, the World Wide Web Consortium (W3C), the international organisation which sets standards for the Web and how it operates, produced a recommendation for three profiles of the Web Ontology Language OWL 2 [A]. One of these, **OWL 2 QL**, is intended for use in VKG systems and its definition cites Kontchakov and Zakharyaschev's work [1] as one of the two main standard citations in the definition of the OWL 2 QL profile [A].

Three major systems using OWL 2 QL to define their VKGs have been developed since the language was standardised: **Mastro**, **Ultrawrap**, and **Ontop**. The businesses founded on these systems have their own economic impact as well as supporting influential projects with commercial and industrial partners. All of these use OWL 2 QL to define their VKGs and thus draw directly on Kontchakov and Zakharyaschev's research.

The Ontop system, initiated at the Free University of Bozen-Bolzano (UniBZ), is that with which Kontchakov and Zakharyaschev have been most closely involved, making key contributions to its development [B] through the collaboration with UniBZ that began in 2007. Their research has been fundamental both to Ontop's theoretical underpinnings [2, 3, 4] and to the code base itself: Kontchakov has made more than 2,000 commits to the code base, the second-highest number among all contributors [B]. Ontop is available open-source on Github (github.com/ontop/ontop) and has become one of the leading Virtual Knowledge Graph systems worldwide [B]. Since 2015, it has been bundled with downloads of Stanford

University's Protégé, an ontology development platform with over 366,000 users (Dec 2020).

In April 2019, UniBZ spun out the Ontop work into a start-up company, **Ontopic s.r.l.**, which now employs three full-time staff who work alongside UniBZ academics to develop tailored commercial solutions based on the Ontop framework. In addition to the economic benefit to those employed, the development of spin-out companies such as Ontopic serves UniBZ's goals as an institution: 'technology and knowledge transfer is the third pillar of the university... joint projects ensure the practical relevance of research and education'. (www.unibz.it/en/home/companies-and-partnerships/knowledge-technology-transfer).

Other VKG systems follow a similar pattern of development. Mastro, built at Sapienza Universita di Roma (Italy), was the first of the three. In 2017, it was spun out as **OBDA Systems s.r.l.**, a commercial start-up with 12 staff (Jan 2021). Similarly, Ultrawrap was developed at the University of Texas, Austin, and commercially spun out in 2015 as **Capsenta**, raising venture capital funding of US$210,000. In 2019 it was acquired by data.world.

All three of these companies work directly with clients in a wide variety of industrial and commercial sectors, employing VKG technology to make complex data easier to access; creating cost savings, opening up inaccessible data repositories, and enabling workers to make better-informed decisions more quickly. Some notable use cases are described below.

Energy sector: In 2012-16, the European Union granted FP7 research and innovation framework funding worth €9 million to the **Optique** Integrated Project (optique-project.eu). This was supplemented by an additional €5 million from industrial partners. Optique developed VKG technology based on the Ontop platform for applications in the energy sector. Researchers worked with Norwegian multinational energy firm Equinor (formerly Statoil) and with German industrial manufacturing conglomerate Siemens to develop exemplar VKG tools, in which Ontop was a core component [B]. The Equinor tool brought together the various data used by exploration geologists in a unified system that reduced query times (which had previously required specialist IT intervention) from several weeks to just a few minutes [C]. These VKG artefacts were subsequently developed (with direct contributions from Kontchakov and Zakharyaschev) and released as a publicly-available VKG specification [E].

The proof-of-concept research carried out under Optique led directly to the development of the **SIRIUS Centre for Scalable Data Access in the Oil and Gas Domain**. 'Build[ing] on the successful formula of Optique' [D, 2016], the Norwegian government established SIRIUS in 2015, in collaboration with industrial partners including Kadme, Evry, Equinor, Bosch, Schlumberger and IBM. This centre, now halfway through its scheduled eight-year research programme, has to date (Dec 2020) received a total of NOK112,885,000 (approximately £9.5 million) in funding, creating new partnerships between academia and industry, and placing VKG technology at the heart of the Norwegian energy industry's ongoing development [D].

Medical: VKGs have numerous applications in clinical settings, where practitioners are typically required to evaluate large quantities of data which may be dispersed across disparate locations. In Brazil, Ontop forms the basis for a VKG system named **Recruit**, implemented at the AC Camargo Cancer Centre in São Paulo in 2014. Founded in 1953, AC Camargo is a major specialist centre for the disease, treating approximately 14,000 patients annually in a total of 950,000 visits and appointments. Its research and learning centre, CIPE, employs 62 research support professionals and is currently supporting approximately 300 ongoing research projects, including clinical trials [J]. Recruit was designed to streamline the participant recruitment for these projects. Before this point, researchers seeking participants had to navigate a diverse information sources, and searches for suitable participants were time-consuming and required technical expertise. Recruit has addressed this problem. Bringing together detailed information on more than 500,000 patients, it allows researchers to search for suitable participants themselves, without an IT intermediary, making it quick and easy for researchers to find detailed information about participants whose clinical criteria marks them as potential candidates [F, I]. Since the system was introduced in 2014, over 400 researchers have made over 160,000 searches [I]. Beneficiaries include the researchers, medical practitioners and nurses working at the hospital; the technical staff who no longer have to routinely handle queries of this nature but can instead devote their attention to developing and improving the centre's software tools; the thousands of patients who are served by the AC

Camargo; and the millions worldwide who benefit from the research carried out there.

Transport and infrastructure: VKGs have proven useful in providing access to open data repositories which facilitate the smoother running of regional infrastructure. Ontopic's ongoing projects in this sector include an €80,000 collaboration with the Italian province of **South Tyrol** to extend their tourism open data portal [B]. Tourism is a key economic driver for the area, which in 2017/18 recorded over 33 million overnight stays from visitors. So far, applications using the data include the region's main tourist website and its associated mobile app, and public transport, traffic, and parking apps for cities within the region. All of these improve quality of life for residents and visitors.

In 2019, Sapienza University used the Mastro system to develop an ontology for the **Automobile Club d'Italia**, mapping its public vehicle register and car tax domains. The result is a publicly available database of Italian car ownership from 2017 onwards, allowing analysis at the national and regional level of information on the 2 million vehicles sold within the country each year and supporting the development of this major market sector (lod.aci.it).

Education/innovation: Founded in 2010, **SIRIS Academic** is a consultancy and think-tank based in Barcelona which employs over 30 staff and specialises in 'semantic modelling, knowledge management and knowledge transfer' for the European higher education and research sector. Since 2014, SIRIS has drawn on Ontop's VKG technology to provide information solutions for its clients, describing Ontop as indispensable to its work [J].

SIRIS's initial work with Ontop came in the context of EPNet, a €2,400,000 ERC-funded project that integrated three Roman archaeological databases into a user-friendly interface allowing scholars to easily run searches across them. Since then, SIRIS has expanded this work to become a central part of its business offering. Notably, Ontop underpins SIRIS's UNiCS (unics.cloud), an Open Data platform based on semantic technologies that integrates an ever-growing number of repositories and datasets about the higher education, research and innovation sector in Europe [J]. Approximately 10,000 users each year from institutions including universities, local governments, and regional agencies responsible for research and development [J] use UNiCS (and the customised portals built from it) to better understand their operating context, allowing them to make informed strategic decisions for the future.

SIRIS also uses UNiCS as the basis for customised data mining applications and strategic solutions for its clients. Commissioned by the **Regional Council of Tuscany**'s Conference for Research and Innovation, the Toscana Open Research Portal (www.toscanaopenresearch.it) integrates data from the Italian central government with other EU data sources to present an information dashboard on the research ecosystem within the region. This is used by universities and regional authorities to generate content and debate on the research and development ecosystem as well as to address strategic needs [J]; for example, during the coronavirus pandemic the portal was used to map regional competencies in digital health, allowing the region better to coordinate its response. The Italian Ministry of Research and Education is now exploring the possibility of scaling up the portal to the national level [J].

In France, after a merger with [redacted] in 2018, the [redacted university] sought SIRIS' services to augment its data analytics capabilities. SIRIS has been using semantic web technology to build tools which allow for easier and more useful analysis of the university's existing data on its operations, much of which was held across multiple databases and structured in such a way as to make long-term comparisons and big-picture planning very difficult to carry out. Using Ontop allows SIRIS to reconfigure the data in ways that are most useful for analysis without materially restructuring its original form. The tools are used by high-level decision-makers within the university to make strategic decisions about partnerships, specialisation and student recruitment [J].

Politics and government: On average, five countries each year draft a new constitution, and 30 update their existing documents. Examining existing constitutions can help to guide the drafting of new ones in terms of wording and scope. However, historically such constitutions were distributed across over numerous databases worldwide, making search time-consuming and in some cases impossible. In the **Constitute** project, launched at the General Assembly of the United Nations in 2013, Ultrawrap was used to integrate the world's constitutions (held in over 195 different databases) into a single unified endpoint for contextual searching. This

makes these vital documents easily accessible to those drafting or redrafting national constitutions, as well as to other stakeholders such as UN agencies and NGOs. Since publication, Constitute has been an 'indispensable resource' for NGOs and UN agencies, informing constitution building in countries including Myanmar, Nepal, South Sudan, Cuba and The Gambia (constituteproject.org/content/testimonials).

As well as the three systems built around OWL 2 QL (**Mastro**, **Ultrawrap**, and **Ontop)**, Kontchakov and Zakharyaschev's work has been instrumental in the development of two other knowledge graph systems, Stardog and Ontotext**.**

Stardog was founded in 2005 by researchers from the University of Maryland's AI lab and is now run through a private company. It has received a total of US$23,300,000 (Dec 2020) funding from private investors (crunchbase.com/organization/Stardog-Union); and employs a staff of over 60 people (stardog.com/about) to work with a range of clients including NASA, eBay, Nokia, and the US Department of Defence. Since December 2015, one of Stardog's 'major value propositions' [H] has been the ability to integrate data stored in relational databases. Stardog used the Ontop framework to provide this service from launch until May 2017, when it was replaced with a new implementation. This was a period of major expansion, during which Stardog reconfigured its charging structure and increased its monthly recurring revenue by 4,500% [H]. Over that time, [Ontop] proved its value for [Stardog] by enabling its customers to integrate numerous databases into a single, coherent Virtual RDF graph [G]. Using Ontop enabled Stardog to enter the market at the most opportune moment and develop this area of their business without the delay that developing their own software entailed.

Ontotext is a Bulgarian company best known for GraphDB, a KG platform that allows users to link data, index it for searching and enrich it using text analysis. Having focused on materialised KGs, Ontotext has become increasingly interested in the possibilities presented by VKGs and in 2020 they integrated the Ontop framework into GraphDB to provide virtualisation of KGs. As well as supporting the implementation of specific use cases, Ontop helps Ontotext to serve its customers by facilitating the swift development of minimum viable graphs (that is, graphs which contain the minimum data necessary to fulfil their function); clients expect these to be delivered quickly. Ontop's adherence to public standards like R2RML is also helpful in reassuring Ontotext's client base about the vendor risk associated with working with Ontotext, something that can be a key barrier for small technology businesses. These functions will become increasingly essential as the market expands: Ontotext anticipates a 40% revenue growth in the sector over the next three years [K].

5. Sources to corroborate the impact (indicative maximum of 10 references)

A. W3C standard for OWL 2 QL, published at www.w3.org/TR/owl2-profiles
B. Testimonial: Dr Benjamin Cogrel, President and CEO of Ontopic, s.r.l.
C. E. Kharlamov, M. Skjaeveland, D. Hovland, T. Mailis, E. Jimenez-Ruiz, G. Xiao, A. Soylu, I. Horrocks, A. Waaler. Finding Data Should be Easier than Finding Oil. In *IEEE International Conference on Big Data*, 2018. doi.org/10.1109/BigData.2018.8622035
D. SIRIUS annual reports, 2016-2020: sirius-labs.no/results-downloads
E. Testimonial, Director of SIRIUS
F. D. F. C. Patrão, M. Oleynik, F. Massicano, A. M. Sasso. 'Recruit - An Ontology Based Information Retrieval System for Clinical Trials Recruitment.' doi.org/10.3233/978-1-61499-564-7-534
G. Testimonial: VP of Research and Development, Stardog
H. Kendall Clark, 'Reviewing 2016, Previewing 2017', Stardog blog: www.stardog.com/blog/reviewing-2016-previewing-2017/
I. Testimonial: Medical Informatics Analyst, AC Camargo Cancer Centre
J. Testimonial: Director of Strategy & Business Development, SIRIS Academic
K. Testimonial: Chief Technical Officer, Ontotext