Institution: University of Oxford

Unit of Assessment: 11 Computer Science and Informatics

Title of case study: Enabling Applications of Ontologies via Reasoning Systems

Period when the underpinning research was undertaken: 2008 – 2019

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
Ian Horrocks	Professor of Computer Science	Sep 2007 – present
Boris Motik	Professor of Computer Science	Sep 2007 – present
Bernardo Cuenca Grau	Professor of Computer Science	Sep 2007 – present
Yavor Nenov	RA/SRA	Oct 2011 – Nov 2019
Robert Piro	RA	Oct 2011 – Jun 2017

Period when the claimed impact occurred: 1 August 2013 – 31 July 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

Research led by Horrocks, Motik, and Cuenca Grau in the Oxford Knowledge Representation and Reasoning group has had a wide-ranging impact on the development and deployment of semantic technologies over the last decade. They played a leading role in developing the revised and extended World Wide Web Consortium (W3C) standard Web Ontology Language, OWL 2, and have developed state-of-the-art reasoning systems that support this language. These represent important advances in exploiting the potential of semantic technologies for complex data and knowledge applications, and the work has led to diverse impacts across both industry and public service settings within the period. Open source reasoning tools developed in the group are enabling applications of OWL ontologies in areas as diverse as global healthcare IT (in the SNOMED clinical terminology) and large-scale infrastructure design (at the Norwegian multinational Aibel). The researchers have also developed RDFox, a high-performance knowledge graph and semantic reasoning engine that has created direct economic impact through the spinout company Oxford Semantic Technologies (OST). OST is backed by Samsung Ventures, with customers including Samsung and Festo, a German multinational industrial control and automation company.

2. Underpinning research

The OWL 2 ontology language and the Resource Description Framework (RDF) form part of the W3C's Semantic Web stack. These standardised formats have been key enablers for tool development and the deployment of semantic technologies across a wide range of domains and applications. At Oxford, Cuenca Grau, Horrocks, and Motik have developed novel algorithms and pioneered state-of-the-art reasoning systems that support these standards, and developed technology for reasoning over RDF data and OWL ontologies that efficiently exploits the greater flexibility afforded by knowledge graphs over traditional database systems.

Reasoning systems are crucial for the application of ontology-based systems, underpinning both ontology development and ontology-based data access. HermiT is currently the only reasoner that fully supports the OWL 2 standard. It supports all of the datatypes specified in the standard, and correctly reasons about properties (binary predicates) as well as about classes (unary predicates). Building on initial work done at Manchester, Horrocks and Motik developed the system at Oxford from 2007, supported by the EPSRC standard grant "HermiT" (2008–12). HermiT uses a novel hypertableau algorithm that generalises so-called absorption optimisations and provides greatly improved efficiency [**R1**]. The system also employs a wide range of novel optimisation techniques, in particular techniques for reducing the size of the constructed models, and supports several extensions that go beyond the OWL standard [**R2**, with international collaborators].

ELK is a consequence-based reasoner developed at Oxford. ELK uses novel algorithms that extend the range of consequence-based reasoning to include more expressive ontology languages that go beyond the OWL 2 EL profile, a standardised subset of OWL that enjoys polynomial time worst-case complexity for many reasoning tasks. A key feature of these algorithms is that, although the language that they support is not worst-case polynomial, they

remain optimal for subsets that do enjoy this property, as proven via parameterised complexity analysis [**R3**]. ELK also employs novel parallelisation techniques to improve efficiency on modern multi-processor architectures.

RDF is a standard model for data exchange on the internet and a key tool for semantic technologies. It extends the linking structure of the Web to use Universal Resource Identifiers (URIs) to name the relationship between things as well as the two ends of the link (a "triple"). This simple model allows structured and semi-structured data to be mixed, exposed, and shared across different applications. The main focus of HermiT and ELK is ontology reasoning: they do support basic reasoning with RDF data, but not conjunctive queries (the basis for the W3C standard query language SPARQL), or scalability to large datasets. RDFox is a highly optimised RDF-triple store developed at Oxford by Motik, Nenov, Piro, and Horrocks. RDFox uses a novel main memory architecture and parallelisation techniques to support storage and efficient query answering over billions of triples [**R4**]. Query answering takes into account not only the RDF data, but also an ontology expressed using the OWL 2 RL profile, which can be seen as a fragment of Datalog. In order to make this efficient, RDFox pre-materialises relevant entailments, and employs novel incremental materialisation algorithms to avoid costly recomputation when the underlying RDF data changes [**R5**]. This is particularly difficult when combined with reasoning about equality, which is required in order to support the OWL and SPARQL standards [**R6**].

3. References to the research

[**R1**] B. Motik, R. Shearer, I. Horrocks: Hypertableau Reasoning for Description Logics. J. Artif. Intell. Res., 2009: https://doi.org/10.1613/jair.2811. *Submitted to REF 2014.*

[**R2**] B. Glimm, I. Horrocks, B. Motik, G. Stoilos, Z. Wang: HermiT: An OWL 2 Reasoner. J. Autom. Reasoning, 2014: https://doi.org/10.1007/s10817-014-9305-1. *Submitted to REF 2021.*

[**R3**] F. Simancik, B. Motik, I. Horrocks: Consequence-based and fixed-parameter tractable reasoning in description logics. Artif. Intell., 2014: https://doi.org/10.1016/j.artint.2014.01.002. *Submitted to REF 2021.*

[**R4**] B. Motik, Y. Nenov, R. Piro, I. Horrocks, D. Olteanu: Parallel Materialisation of Datalog Programs in Centralised, Main-Memory RDF Systems. Proc. of the 28th Nat. Conf. on Artif. Intell. (AAAI), 2014: https://ojs.aaai.org/index.php/AAAI/article/view/8730.

[**R5**] B. Motik, Y. Nenov, R. Piro, I. Horrocks: Maintenance of datalog materialisations revisited. Artif. Intell., 2019: https://doi.org/10.1016/j.artint.2018.12.004. *Submitted to REF 2021.*

[**R6**] B. Motik, Y. Nenov, R. Piro, I. Horrocks: Combining Rewriting and Incremental Materialisation Maintenance for Datalog Programs with Equality. Proc. of the 24th Int. Joint Conf. on Artif. Intell. (IJCAI), 2015: http://ijcai.org/Abstract/15/441.

Grants. The research was supported by a series of EPSRC grants awarded to Horrocks, Motik, and Cuenca Grau between 2008 and 2021 and totalling around GBP4,300,000 (HermiT, EP/F065841/1; ConDOR, EP/G02085X/1; DBOnto, EP/L012138/1; ED3, EP/N014359/1; MaSI3, EP/K00607X/1; AnaLOG, EP/P025943/1).

4. Details of the impact

Our work on reasoning infrastructure for ontologies and knowledge graphs has led to a range of impacts in diverse settings. These include:

- impact on ontology developers, through the development of technical standards and open source tools that support these standards;
- impact on non-profit organisations and businesses that develop and use ontology-based systems;
- direct impacts on commerce and production through the spinout company OST and the application of its patented technology in industry.

Impact on ontology development through open source reasoners for OWL 2 ontologies.

Horrocks, Motik, and Cuenca Grau have been influential in the development of the OWL ontology language, editing key specification documents for the refined and extended OWL 2 standard (second edition published by W3C in 2012) [**E1**]. The standard allows ontologies to be shared across applications, and has stimulated increasing use of ontologies in a range of sectors. Reasoning systems that support OWL, in turn, are an important enabling factor for the wider adoption of the standard and the increased uptake of ontology-based systems. Effective reasoners are critical for ontology development and use, for example to verify the consistency of the knowledge they express, to make implicit knowledge explicit, or to access data via queries. The OWL reasoners HermiT [**R1**–**2**] and ELK [**R3**] have achieved wide reach as open source projects and through distribution with the open source Protégé OWL ontology editor, in which HermiT – the only reasoner that fully and correctly supports OWL [**R2**] – is the standard reasoner. Protégé is the most popular tool for ontology design, with over 360,000 registered users, and is widely relied on for building and maintaining ontologies in industry and in large government projects such as the National Cancer Institute Thesaurus and the WHO's International Classification of Diseases [**E2**].

Large clinical and biomedical nomenclatures like these are a prominent application area for ontologies; they are gradually superseding existing medical classifications, and will provide the future platforms for gathering and sharing medical knowledge. One such ontology is SNOMED CT, the world's most comprehensive clinical terminology – "a codified vocabulary that is now accepted as a common global language for health terms" [**E3**]. It is owned and administered by the non-profit organisation SNOMED International, and is officially used in over 80 countries (including the UK, a founding charter member) to support the direct management of individual health and care and to improve the flow of data across health and care systems. SNOMED CT is an NHS Digital information standard, applying to all organisations providing NHS, Public Health, and/or Adult Social Care services. Since 2018 it has been implemented across Primary Care settings within the NHS, and is a core component of the NHS patient record service; all secondary, dentistry, and optometry services are required to establish a detailed implementation schedule by 31 December 2020 [**E3**, **E4**].

SNOMED International has used ELK since 2016 to perform complex reasoning tasks in its new authoring platform: "for example, to check the logical consistency of (the definitions of) SNOMED concepts, and to automatically organise concepts into a hierarchical structure, which is essential when dealing with such a large terminology" [**E3**]. SNOMED CT contains more than 400,000 terms and is "under constant revision and extension". "ELK was a game changer for us as it reduced the time needed to reason with SNOMED CT from hours/days to only a few seconds. Not only that but, unlike some earlier systems, ELK was developed by researchers who were also able to provide formal and peer-reviewed guarantees as to the correctness of the algorithms embodied in ELK [**R3**]. This is clearly of great value given the critical role played by SNOMED CT in the NHS and many other national healthcare systems" [CIO, SNOMED International: **E3**].

B2i Healthcare is a software engineering firm that specialises in SNOMED CT and healthcare information standards and exchange. B2i has used ELK across the whole assessment period to provide reasoning services in its authoring platform, Snow Owl. Snow Owl is deployed in over 3,000 locations in over 80 countries, and B2i is able to state that "ELK is the de facto global standard for classifying SNOMED CT ontologies by national eHealth programs and organizations". These include national health ministries and agencies in the UK, Belgium, Denmark, Estonia, Ireland, Norway, Sweden, Switzerland, Australia, Singapore, New Zealand, and the USA [Non-Executive Chairman, B2i Healthcare: **E5**]. B2i uses ELK because of "its outstanding performance and the correctness guarantees provided by its basis in world leading research [**R3**]. ELK performs description logic classification in parallel on modern multi-core computers, which allows the full international SNOMED CT plus the Australian extensions (830,926 relationships) to be classified and checked for equivalencies in about 10 seconds on a modern desktop computer. This is in stark contrast to earlier systems which took hours to perform this task, if they were capable of performing it at all. The orders of magnitude

performance improvement provided by ELK means that reasoning can be used 'on the fly' during the authoring process, resulting in both improved quality and productivity" [**E5**].

Besides medical and life sciences, OWL ontologies are increasingly used in industry to address information integration and access problems. Aibel is a multinational company headquartered in Norway that provides engineering and construction services in the oil, gas, and offshore wind industries. Aibel has used HermiT across the whole assessment period to (i) support the development of a new master ontology system for design requirements and specifications; and (ii) perform reasoning and query answering over this custom-built ontology to configure design and product selection. These tasks are now "performed with greater precision and less effort than with Aibel's legacy system, ultimately resulting in a design of higher quality, which again reduces the total time and cost of construction" [Senior Manager: **E6**]. Aibel uses HermiT because of "its full support for the OWL ontology language, and its superior performance, reliability and scalability" in ontology reasoning [**E6**; see **R1**–**R2**]. Using HermiT, Aibel has been able to document significant reductions in data errors and duplications. In the context of the very large projects delivered by the company, such deficiencies in data quality can have significant cost implications: for example, the presence of duplicate design artefacts leads to erroneous bulk orders of materials. The knowledge-based approach enabled by HermiT supports "a more efficient and precise description of design artefacts…removing practically all duplicate design artefacts recorded in the system. (It is estimated that more than 30% of the legacy system's data was duplicate data.) The lack of duplicates and added detail in design descriptions make it easier to manage material storage and order a better selection of materials…The effect is an estimated cost reduction of approximately 5% for bulk material orders, which in large projects amounts to more than EUR100,000,000" [**E6**]. Aibel states that: "in summary, HermiT plays an important role at Aibel, and has enabled us to deploy knowledge-based solutions that significantly reduce cost" [**E6**].

Impact in industry through commercialisation of high-performance RDFox system.

Pathway to impact. The highly scalable RDFox tool was initially released for research purposes in 2014 under an open source licence. A patent was also filed on the underpinning research: this would later be significant to the commercialisation of the technology through the spinout company OST [**E7**]. The Oxford team subsequently worked with industrial partners to develop the technology into a mature platform with performance bottlenecks identified and removed [**R5**, **R6**]. RDFox was applied in several pilot industry projects, including with Siemens, for ontology-based data access enhancement (2015); Statoil (now Equinor), for integration and analysis of oil production and geological survey data (2016); Kaiser Permanente, in patient data analysis for healthcare compliance (2016); EDF Energy, to manage and analyse information about their electricity distribution network (2016); and other partners including Skyscanner, for recommender engines, and Armasuisse for text analysis [**E8**]. During these pilot projects the functionality of RDFox was extended to match business requirements.

In 2017, OST was formed to bring cutting-edge research in semantic technologies to industry. OST has raised GBP4,100,000 in investment, including GBP3,000,000 in Series A investment led by Samsung Ventures, announced in June 2019 [**E9**]. By September 2019 OST employed 6 full-time staff; 9 FTEs were employed by the end of the assessment period. OST's patented technology is sold under licence to customers for a fee of approximately GBP50,000 per licence. Since April 2018, the company has secured licence sales worth over GBP570,000 [**E10**]. Customers include Festo, a German multinational production line equipment company, and electronics giant Samsung.

Festo produce and sell pneumatic and electrical control and drive technology for factory or process automation. They have used RDFox since 2018 to implement a new semantic approach to data processing through the Festo Semantic Platform [**E11**, **E12**.2]. In turn, the semantic approach has allowed them to optimise their sales process. Festo's "choice of RDFox was based on its outstanding performance, reliability and scalability" compared to other knowledge graph systems, and its basis "in published algorithms whose correctness has been formally verified", giving "high confidence that RDFox will continue to provide correct answers regardless of the knowledge and the queries that we use" [Head of Smart Data Services: **E11**].

RDFox's performance advantages have a direct impact on production at Festo. The company's product catalogue includes thousands of components that can be combined in millions of different configurations, with configurations being subject to many complex constraints. RDFox is used to compute possible configurations and to ensure their validity. The previous approach, using multiple relational databases, "took several hours to compute valid configurations, and also required complex data maintenance and deployment management. In contrast, RDFox can compute valid configurations in a few seconds…This saves us a huge amount of time and also allows for much greater flexibility in updating our product catalogue" [**E11**]. The company has found that "the semantic approach is also much easier to maintain and extend" [**E11**]. Queries are now "much easier to write", and "extensions or changes in the information model due to technical or business demands are much faster accomplished and validated. That is a significant increase in flexibility as well as saving time and computing resources" [**E12**.1]. As a result, Festo are extending the Semantic Platform to cover further technical product domains. As "a key component of the Festo Semantic Platform", RDFox is central to this new approach across the company, "and as such is already having a significant impact on our data management business" [**E11**; see also **E12**.2].

Prior to their investment in OST, Samsung had already funded various projects, worth approximately GBP500,000, with a view to using RDFox in several different customer-facing applications. Samsung are now using the RDFox knowledge graph system "to drive applications running on a variety of platforms including embedded systems and mobile devices (phones and tablets)"; these include, for example, lifestyle recommender systems linking very large and heterogeneous data sets covering areas such as food, health, and exercise. "When fully deployed, [Samsung] expect such applications to be running on millions of devices, all of which will access relevant knowledge via RDFox" [**E13**]. Samsung's choice of RDFox "was based on its outstanding performance, reliability and scalability." Other knowledge graph systems tested by Samsung "were all found to be much slower than RDFox, often by orders of magnitude, and could also produce incorrect answers…In summary, RDFox is a key component of our knowledge-intensive applications, and we expect such applications to be used by hundreds of millions of Samsung customers worldwide" [**E13**].

5. Sources to corroborate the impact

[**E1**] OWL 2 Web Ontology Language. Document Overview (2nd edition; editor details at foot of the document): https://www.w3.org/TR/2012/REC-owl2-overview-20121211/. OWL Working Group (closed 2012): https://www.w3.org/2007/OWL/wiki/OWL_Working_Group.
[**E2**] Protégé open source ontology editor (https://protege.stanford.edu/): (1) user numbers; (2) documentation, showing Hermit is built-in reasoner; (3) use in NCI Thesaurus and WHO ICD.
[**E3**] Letter of Aug. 2020 from CIO, SNOMED International, on the impact of ELK.
[**E4**] SNOMED CT in the NHS. Overview: https://tinyurl.com/yy7g4ttz; NHS information standard SCCI0034 (SNOMED CT): https://tinyurl.com/sgj635y.
[**E5**] Letter of Aug. 2020 from Non-Executive Chairman, B2i Healthcare, on the impact of ELK.
[**E6**] Letter of Oct. 2020 from Senior Manager, Aibel, on the impact of HermiT.
[**E7**] US patent US 20160259796 A1 (filed Oct 2014, granted Oct 2020): http://tiny.cc/xf11tz.
[**E8**] Papers reporting technology-evaluation projects with industry partners (DOI/URL): Siemens – 10.1007/978-3-319-11964-9_38; 10.1145/2933267.2933290; Equinor – 10.1007/978-3-319-25010-6_6; Kaiser Permanente – 10.1007/978-3-319-46547-0_34; EDF Energy – http://tiny.cc/3kn0tz.
[**E9**] Samsung investment in OST: https://innovation.ox.ac.uk/news/samsung-invests-ost/.
[**E10**] Letter of Jan. 2020 from CEO at OST, corroborating information about the company.
[**E11**] Letter of Sep. 2020 from Head of Smart Data Services, Festo, on the impact of RDFox.
[**E12**] Industry papers on RDFox at Festo. (1) Paper by Festo on their application of RDFox in factory automation: https://iswc2017.semanticweb.org/paper-462/; (2) Paper by Festo on RDFox and the Festo Semantic Platform: https://doi.org/10.1007/978-3-030-32327-1_9.
[**E13**] Letter of Aug. 2020 from Head of Bigdata Team, Samsung Research, on use of RDFox.