

Journal of Zhejiang University-SCIENCE C (Computers & Electronics)
 ISSN 1869-1951 (Print); ISSN 1869-196X (Online)
 www.zju.edu.cn/jzus; www.springerlink.com
 E-mail: jzus@zju.edu.cn



Perspective:

Semantics \sqcap scalability $\models \perp$?

Ian HORROCKS

Department of Computer Science, Oxford University,
 Oxford, UK

E-mail: Ian.Horrocks@comlab.ox.ac.uk

doi:10.1631/jzus.C1101001

The Web Ontology Language (OWL) (Horrocks *et al.*, 2003; W3C OWL Working Group, 2009a) has been developed and standardised by the World Wide Web Consortium (W3C). It is one of the key technologies underpinning the Semantic Web, but its success has now spread far beyond the Web: it has become the ontology language of choice for applications in fields as diverse as biology (Sidhu *et al.*, 2005), medicine (Golbreich *et al.*, 2006), geography (Goodwin, 2005), astronomy (Derriere *et al.*, 2006), agriculture (Soergel *et al.*, 2004), and defence (Lacy *et al.*, 2005). Moreover, ontologies are increasingly being used for ‘semantic data management’, and DB (database) technology vendors have already started to augment their existing software with ontological reasoning. For example, Oracle Inc. has recently enhanced its well-known database management system with modules that use ontologies to support ‘semantic data management’. Their product brochure (http://www.oracle.com/technology/tech/semantic_technologies/pdf/oracle%20db%20semantics%20overview%2020080722.pdf) lists numerous application areas that can benefit from this technology, including enterprise information integration, knowledge mining, finance, compliance management, and life science research.

The standardisation of OWL has brought with it many benefits. In the first place, OWL’s basis in description logic has made it possible to exploit the results of more than 25 years of research and to directly transfer theoretical results and technologies to OWL. As a consequence, algorithms for computing OWL entailment are well known (Haarslev *et al.*, 2004; Tsarkov and Horrocks, 2006; Sirin *et al.*, 2007; Motik *et al.*, 2009), and the formal properties of the problem are well understood: it is known to be decidable, but to have

high complexity (NExpTime-complete for OWL and 2NExpTime-complete for OWL 2 (W3C OWL Working Group, 2009b)).

Entailment is a very general reasoning task to which many other kinds of reasoning can be reduced. It is common to distinguish two general categories of reasoning tasks: those that are concerned primarily with classes, and those that are concerned primarily with individuals. Common tasks in the first category include checking class (un-)satisfiability (a class C is unsatisfiable with respect to (w.r.t.) an ontology \mathcal{O} iff $\mathcal{O} \models C \equiv \perp$), subsumption (a class C subsumes a class D w.r.t. \mathcal{O} iff $\mathcal{O} \models D \sqsubseteq C$), and classification (computing the subclass quasi-order for all the class names occurring in an ontology \mathcal{O}) (Baader *et al.*, 2008). The most common task in the second category is query answering (given an ontology \mathcal{O} and a query q , compute the set of tuples of individuals $\{\bar{a} \mid \mathcal{O} \models q(\bar{a})\}$) (Horrocks and Tessaris, 2002).

1 OWL reasoning

Notwithstanding the high worst case complexity of the underlying problem, highly optimised reasoners for class reasoning in OWL, including, e.g., ELK (Kazakov *et al.*, 2011), FaCT++ (Tsarkov and Horrocks, 2006), HermiT (Motik *et al.*, 2009), Pellet (Sirin *et al.*, 2007), and Racer (Haarslev and Möller, 2001), are now extremely effective in practice (Tables 1 and 2). There has also been significant progress on improving the performance of such systems with respect to query answering. However, inherent limitations of the underlying algorithms, and the typically large size of data sets, mean that many realistic query answering problems are still too hard for such systems.

A range of different algorithmic approaches have been developed in an effort to address this issue, each with its own strengths and weaknesses. We will briefly review the most prominent approaches.

1.1 Query rewriting

In an effort to improve scalability, systems such as Mastro (Calvanese *et al.*, 2007b), Quonto

Table 1 Performance of OWL reasoners on large ontologies

Ontology	Logic	Number of classes	Number of properties	Number of axioms	Number of subsumptions	ELK*	HermiT	Pellet	PacT++
SNOMED CT	\mathcal{EL}	315 489	58	430 844	$> 10^{11}$	(13.15, 5.02)	–	–	–
GALEN	\mathcal{EL}	23 136	950	36 547	$> 10^8$	(1.33, 0.77)	–	–	–
FMA	\mathcal{EL}	78 977	7	121 712	$> 10^9$	(0.44, 0.39)	–	–	–
GO	\mathcal{EL}	19 468	1	28 897	$> 10^8$	(0.20, 0.19)	–	–	–
Plant Anat.	\mathcal{SHIF}	19 145	82	35 770	$> 10^8$	(–, –)	11.2	87.2	22.9
SWEET-P	\mathcal{SHOIN}	1728	145	2419	$> 10^6$	(–, –)	11.2	–	0.2
NCI-2	\mathcal{ALCH}	70 576	189	100 304	$> 10^9$	(–, –)	–	172.0	60.7
DOLCE-P	\mathcal{SHOIN}	118	264	265	$> 10^4$	(–, –)	105.1	105.1	–

*(1 worker, 4 workers)

Table 2 Evolution of reasoner performance over time

Year	Ontology size	Complete reasoning	Classification time (s)
1995	3000	No	10^5
1998	3000	Yes	300
2005	30 000	Yes	30
2010	400 000	Yes	5

(www.dis.uniroma1.it/~quonto), and Requiem (www.comlab.ox.ac.uk/projects/requiem) employ a query rewriting approach in which the ontology is used to rewrite the query into a union of conjunctive queries (UCQ) whose evaluation can be delegated to a standard database system (Calvanese *et al.*, 2007a). Unfortunately, the ontology language must be quite restricted in order to guarantee that such a rewriting exists (Artale *et al.*, 2009), and such systems can support only the OWL 2 QL subset of OWL (W3C OWL Working Group, 2009b). Using OWL 2 QL it is possible to model, for example, class hierarchies and (some) incomplete information, but not disjunctive information, relationship cardinalities, or property chaining. Moreover, even with very restricted ontology languages, query rewriting algorithms can in theory produce very large rewritings, which cannot be (efficiently) evaluated by standard database systems (Pérez-Urbina *et al.*, 2009), although optimised systems have been shown to work well in practice for OWL 2 QL ontologies (Rodríguez-Muro and Calvanese, 2011).

1.2 Materialisation

Resource Description Framework (RDF) triple stores such as Sesame (www.openrdf.org), OWLim (www.ontotext.com/owlim), Minerva (www.alphaworks.ibm.com/tech/semanticstk), WebPIE (www.few.vu.nl/~jui200/webpie.html), and Oracle's Semantic Data Store (www.oracle.com/technology/tech/semantic_technologies) use a materialisation approach to improve the scalability of query answering. These systems are

based on a relatively loose integration of ontology and database technologies, with the ontology being used in a preprocessing phase to materialise implied facts, after which queries are evaluated over the augmented data and without further reference to the ontology.

Materialisation based systems are now in quite widespread use, and have even been developed to run on mobile platforms such as the Samsung Galaxy II smartphone (Motik *et al.*, 2012). However, although the materialisation technique allows for efficient query evaluation, it also suffers from several drawbacks. Perhaps most important of these is that it can fully support only the OWL 2 RL subset of OWL (W3C OWL Working Group, 2009b). Another drawback is that the augmented data set may be unfeasibly large, and even when it is of 'reasonable' size, it can be very costly to compute. For example, one test data set required 90 h to load into Oracle's Semantic Data Store (www.oracle.com/technology/tech/semantic_technologies/pdf/semtech09.pdf), and although recent work on parallelisation has shown promising results, materialisation of large datasets can still take several hours even on large clusters (Urbani *et al.*, 2012). Finally, the approach is applicable only if the information system fully controls and can modify the data, ruling it out in many practical application settings.

1.3 Combined techniques

More recently, combined techniques have been developed in order to deal with ontology languages that are (slightly) more expressive than those used in pure query rewriting techniques while still allowing query evaluation to be delegated to a data repository (Lutz *et al.*, 2009). Like triple stores, combined techniques use the ontology in a preprocessing phase, but they add only the information that is necessary to allow for subsequent query answering via query rewriting. However, the subset of OWL that can be fully supported is still relatively weak (OWL 2 EL), and excludes important features such as disjunctive information and relationship cardinalities. Moreover, the approach still suffers from some of the

other drawbacks of materialisation techniques: preprocessing may be costly, and the information system must be able to modify the data.

2 Mitigating incompleteness

As discussed above, existing materialisation based procedures can fully handle only a relatively weak subset of OWL, but they can still be used in a sound but possibly incomplete way with an ontology that is outside the relevant subset. In such cases the set of forward chaining inference rules can no longer guarantee to materialise all implied facts, a problem that is inherent to this technique as the materialisation needs to be both deterministic (it must generate a single data set over which queries can subsequently be evaluated) and of bounded size (or it will not terminate). This means that query answers can be incomplete when the ontology contains, e.g., disjunction or existential quantification. In such cases there is usually no way to know if the answer to a query is complete or not, and if incomplete how incomplete.

To address this issue, techniques have recently been investigated for measuring and potentially ‘repairing’ causes of incompleteness. This work has shown that for many ontologies and queries it is possible to identify all relevant causes of incompleteness, and that in many realistic cases it is possible to prove that answers to specific queries are complete even if the query answering system being used is incomplete in general (Cuenca Grau *et al.*, 2011).

As far as ontology repair is concerned, one simple way to eliminate some causes of incompleteness is to use a complete reasoner to add implicit subsumptions to the ontology, an idea that has already been used in the DLDB (Pan and Hefin, 2003) and PelletDB (<http://clarkparsia.com/pelletdb>) systems. This technique is, however, limited to adding simple atomic subsumption axioms, and more sophisticated techniques might make it possible to repair ontologies by adding more complex axioms (Cuenca Grau *et al.*, 2011).

3 Discussion

Reasoning tools are vital for ontology engineering and to support ontology based systems and applications. In the former case, the focus is mainly on class reasoning, and highly effective reasoners are already available. In the latter case, the focus is mainly on query answering, and although great progress has been made, challenges still remain. This is a very active research area, with many different techniques being developed and investigated. Given the size of this research effort, and our ever deepening understanding of both theoretical and prac-

tical issues, it is reasonable to expect that the future performance improvements in query answering systems will be even more spectacular than those achieved in the past by class reasoning systems.

References

- Artale, A., Calvanese, D., Kontchakov, R., Zakharyashev, M., 2009. The DL-Lite family and relatives. *J. Artif. Intell. Res.*, **36**:1-69.
- Baader, F., Horrocks, I., Sattler, U., 2008. Description Logics. *In: Harmelen, F., Lifschitz, V., Porter, B. (Eds.), Handbook of Knowledge Representation. Elsevier, London, p.135-180.*
- Calvanese, D., de Giacomo, G., Lembo, D., Lenzerini, M., Rosati, R., 2007a. Tractable reasoning and efficient query answering in description logics: the DL-Lite family. *J. Autom. Reason.*, **39**(3):385-429. [doi:10.1007/s10817-007-9078-x]
- Calvanese, D., de Giacomo, G., Lembo, D., Lenzerini, M., Poggi, A., Rosati, R., 2007b. MASTRO-I: Efficient Integration of Relational Data Through DL Ontologies. *Proc. 20th Int. Workshop on Description Logics*, **250**:227-234.
- Cuenca Grau, B., Motik, B., Stoilos, G., Horrocks, I., 2011. Completeness guarantees for incomplete ontology reasoners: theory and practice. *J. Artif. Intell. Res.*, in press.
- Derriere, S., Richard, A., Preite-Martinez, A., 2006. An Ontology of Astronomical Object Types for the Virtual Observatory. *Proc. Special Session 3 of the 26th Meeting of the IAU: Virtual Observatory in Action: New Science, New Technology, and Next Generation Facilities.*
- Golbreich, C., Zhang, S., Bodenreider, O., 2006. The foundational model of anatomy in OWL: experience and perspectives. *J. Web Semant.*, **4**(3):181-195. [doi:10.1016/j.websem.2006.05.007]
- Goodwin, J., 2005. Experiences of Using OWL at the Ordnance Survey. *Proc. 1st OWL Experiences and Directions Workshop.*
- Haarslev, V., Möller, R., 2001. RACER system description. *LNCS*, **2083**:701-705. [doi:10.1007/3-540-45744-5_59]
- Haarslev, V., Möller, R., Wessel, M., 2004. Querying the Semantic Web with Racer + nRQL. *Proc. KI-2004 Int. Workshop on Applications of Description Logics.*
- Horrocks, I., Tessaris, S., 2002. Querying the Semantic Web: a formal approach. *LNCS*, **2342**:177-191. [doi:10.1007/3-540-48005-6_15]
- Horrocks, I., Patel-Schneider, P.F., Harmelen, F., 2003. From *SHIQ* and RDF to OWL: the making of a Web Ontology Language. *J. Web Semant.*, **1**(1):7-26. [doi:10.1016/j.websem.2003.07.001]
- Kazakov, Y., Krötzsch, M., Simančík, F., 2011. Concurrent Classification of EL Ontologies. *Proc. 10th Int. Semantic Web Conf.*, p.305-320.
- Lacy, L., Aviles, G., Fraser, K., Gerber, W., Mulvehill, A., Gaskill, R., 2005. Experiences Using OWL in Military Applications. *Proc. 1st Int. Workshop on OWL Experiences and Directions*, **188**.
- Lutz, C., Toman, D., Wolter, F., 2009. Conjunctive Query Answering in the Description Logic EL Using a Relational Database System. *Proc. 21st Int. Joint Conf. on Artificial Intelligence*, p.2070-2075.

- Motik, B., Shearer, R., Horrocks, I., 2009. Hypertableau reasoning for description logics. *J. Artif. Intell. Res.*, **36**:165-228. [doi:10.1613/jair.2811]
- Motik, B., Horrocks, I., Kim, S.M., 2012. Delta-Reasoner: a Semantic Web Reasoner for an Intelligent Mobile Platform. Proc. 21st Int. World Wide Web Conf.
- Pan, Z., Hefflin, J., 2003. DLDB: Extending Relational Databases to Support Semantic Web Queries. Proc. ISWC Workshop on Practical and Scaleable Semantic Web Systems, p.109-113.
- Pérez-Urbina, H., Horrocks, I., Motik, B., 2009. Efficient query answering for OWL 2. *LNCS*, **5823**:489-504. [doi:10.1007/978-3-642-04930-9_31]
- Rodriguez-Muro, M., Calvanese, D., 2011. Dependencies: Making Ontology Based Data Access Work. Proc. 5th Alberto Mendelzon Int. Workshop on Foundations of Data Management, **749**.
- Sidhu, A., Dillon, T., Chang, E., Sidhu, B.S., 2005. Protein Ontology Development Using OWL. Proc. 1st Int. Workshop on OWL Experiences and Directions, **188**.
- Sirin, E., Parsia, B., Cuenca Grau, B., Kalyanpur, A., Katz, Y., 2007. Pellet: a practical OWL-DL reasoner. *J. Web Semant.*, **5**(2):51-53. [doi:10.1016/j.websem.2007.03.004]
- Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., Katz, S., 2004. Reengineering thesauri for new applications: the AGROVOC example. *J. Dig. Inform.*, **4**(4):1-23.
- Tsarkov, D., Horrocks, I., 2006. FaCT++ description logic reasoner: system description. *LNCS*, **4130**:292-297. [doi:10.1007/11814771_26]
- Urbani, J., Kotoulas, S., Maassen, J., Harmelen, F., Bal, H., 2012. WebPIE: a Web-scale parallel inference engine using MapReduce. *J. Web Semant.*, **10**(1):59. [doi:10.1016/j.websem.2011.05.004]
- W3C OWL Working Group, 2009a. OWL 2 Web Ontology Language Overview. Available from <http://www.w3.org/TR/owl2-overview/> [Accessed on Feb. 19, 2012].
- W3C OWL Working Group, 2009b. OWL 2 Web Ontology Language Profiles. Available from <http://www.w3.org/TR/owl2-profiles/> [Accessed on Feb. 19, 2012].

Recommended reading

- Horrocks, I., 2008. Ontologies and the Semantic Web. *Commun. ACM*, **51**(12):58-67. [doi:10.1145/1409360.1409377]
- Krötzsch, M., Simancik, F., Horrocks, I., 2012. A Description Logic Primer. CoRR, abs/1201.4089.
- Baader, F., Horrocks, I., Sattler, U., 2008. Description Logics. In: van Harmelen, F., Lifschitz, V., Porter, B. (Eds.), *Handbook of Knowledge Representation*, Chapter 3. Elsevier, p.135-180.
- Cuenca Grau, B., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U., 2008. OWL 2: the next step for OWL. *J. Web Semant.*, **6**(4):309-322. [doi:10.1016/j.websem.2008.05.001]
- Cuenca Grau, B., Motik, B., Stoilos, G., Horrocks, I., 2011. Completeness guarantees for incomplete ontology reasoners: theory and practice. *J. Artif. Intell. Res.*, in press.

Accepted manuscript available online (unedited version)

<http://www.zju.edu.cn/jzus/inpress.htm>

- As a service to our readers and authors, we are providing the unedited version of accepted manuscripts.
- The section "Articles in Press" contains peer-reviewed, accepted articles to be published in *JZUS (A/B/C)*. When the article is published in *JZUS (A/B/C)*, it will be removed from this section and appear in the published journal issue.
- Please note that although "Articles in Press" do not have all bibliographic details available yet, they can already be cited as follows: Author(s), Article Title, Journal (Year), DOI. For example:
ZHANG, S.Y., WANG, Q.F., WAN, R., XIE, S.G. Changes in bacterial community of anthrance bioremediation in municipal solid waste composting soil. *J. Zhejiang Univ.-Sci. B (Biomed. & Biotechnol.)*, in press (2011). [doi:10.1631/jzus.B1000440]
- Readers can also give comments (Debate/Discuss/Question/Opinion) on their interested articles in press.