From the earliest days of semantic web research, the problem of semantically processing information at very large scale has cast a shadow over its many successes. Having the Web as a primary use case has been both a blessing and a curse. While targeting the Web has attracted a lot of attention and has significantly improved the awareness for the benefits of semantic data models and automatic reasoning, the field still has to prove that semantics will work on web scale. During the first years, semantic web research was dominated by on the consideration of rich semantic representations in the tradition of symbolic AI. Significant progress was achieved with respect to integrating knowledge representation and reasoning with mainstream web infrastructure leading to key standards such as OWL, RDF and SPARQL, however, processing enormous quantities of the corresponding data is still one of the greatest challenges for the Semantic Web. While other communities, e.g. Information Retrieval, have developed successful strategies for coping with the scale of the Web using statistical techniques, semantic web technologies are still struggling with scaling up to the Web as such. This is in part due to the need to preserve the data’s structure and the need to perform various forms of reasoning in order to more effectively leverage the available information. In order to cope with these challenges it is necessary to look beyond the realms of artificial intelligence and to leverage ideas and techniques from the distributed systems and the database community.

There have been a number of successful approaches for combining ideas from these areas with richer semantic representations. Examples include semantically enriched Peer-to-peer systems [1,2], distributed architectures to support reasoning and query processing [3,4,5,6], search engines for semantic data [7,8,9], and novel storage and indexing schemes [10,11]. Recent developments in linked open data has been a driving force for research into the scalability of semantic information processing, because for the first time since the original publications on semantic web technology, the amount of RDF data on the Web has reached a level where traditional RDF stores are pushed to their limits. This special issue features the latest results on combining semantic technologies with technologies from distributed systems and database technologies to push semantic information processing closer to web scale. We have selected four papers out of a number of high quality submissions that represent important topics in the research area.

Karnstedt et al. [12] present an approach for large scale query processing of linked data on the web. The authors apply state of the art database technologies to this problem and adapt existing approaches to the special needs on linked data. The approach is based on the use of distributed DHT-based indexing and tries to exploit distribution and parallelism as much as possible by the introduction of parallel implementations of physical query operators and a strategy for routing subqueries to distributed information stores. The strong point of the work is the combination of best practices in database systems, i.e. the use of algebraic optimizations and a focus on specific problems related to large scale RDF processing, i.e. triple patterns and similarity queries over RDF data. The work therefore contributes to the development of a scalable infrastructure for processing linked open data on web scale using best practices and ideas from database systems.

Delbru et al. [13] present an approach for searching entities on the web using a combination of free-text and semi-structured queries, thereby extending typical information retrieval approaches with a linked data component. The focus of the work is on the creation and maintenance of effective indexing structures for entities. The authors present an intelligent index structure and discuss issues such as incremental updates of the index and compression techniques. The authors also systematically study the performance of different variant of the indexing schema on large scale datasets from the web. The work contributes to the further development of semantic search engines that help users to retrieve structured descriptions of entities instead of documents.
Urbani et al. [14] present WebPIE, an approach to computing the materialization of a large semantic web knowledge base using parallelization. In particular, WebPIE uses the well-known MapReduce programming model. While prior work on parallelization of reasoning has focused on RDF(S), WebPIE supports the more expressive OWL ter Horst language, a fragment of OWL. The authors introduce a number of clever optimizations that enable efficient reasoning. The results are impressive: with 64 nodes, the system computes the closure of 100 billion triples in a blinding 15 minutes.

Hogan et al. [15] present a scalable approach to determine when two identifiers refer to the same entity; sometimes referred to as entity coreference or entity resolution. The authors recognize that Linked Data already has a large number of links and compare applying semantic rules to existing data with more statistically-based similarity approaches. Like WebPIE, parallelization is utilized to achieve scale. One interesting aspect of the work is an attempt to use inconsistencies to detect erroneous coreferences found by the system, so that they can be removed. The authors perform an extensive analysis that considers the messy conditions of real world data.

We conclude that the content of this special issue nicely reflects the current lines of work on large scale semantic information processing. While the work of Karnstedt et al. represents work related to applying database techniques to semantic data, the work of Delbru et al. borrows ideas and principles from the area of information retrieval and the WebPIE system is representative for pushing the borders of logical reasoning further by using distributed systems. Finally, Hogan et al. exploit ideas from distributed systems but also incorporate techniques from outside formal semantics to deal with the inherent noisy nature of the Web.

The work presented here shows that the area of scalable semantic information processing is an active area of research; however it is clear that there is still much to accomplish. One challenge is that work in the relevant areas is still not sufficiently integrated. Although there is some exchange of ideas between the fields, joint work by researchers from AI and other communities is still rare and there are fundamental differences in the any given work will be perceived in the different communities.

A fundamental unsolved issue is the lack of evaluation methods that will convincingly demonstrate web-scalability. While a number of benchmarks have been developed to test scalability of semantic processing methods [16,17], truly proving that semantic systems can achieve web scale is almost infeasible, as it would require a hardware infrastructure that clearly goes beyond the ability of the average research institution. In order to resolve this fundamental problem, the research community must work to find a common agreement on what testable conditions will allow us to extrapolate that a given method can achieve “web scale.”


