In recent years, we have witnessed substantial exploitation of search technologies, both at Web and enterprise scale. However, the representation of user queries and information in existing search appliances is still almost exclusively achieved by simple bag-of-words models (i.e. keyword queries and term-based representation documents). While these systems have shown to work well for many common information needs, they usually fail to address more complex information needs associated with tasks such as aggregation and information analytics.

On the one hand, recent advances in the field of semantic technologies have resulted in tools and standards that allow for the articulation of knowledge at a high level of expressivity. Semantic repositories and reasoning engines have now advanced to a state where querying and processing of this knowledge can scale to large-scale scenarios. As such, semantic technologies are posed to provide significant contributions to Information Retrieval (IR) problems. More expressive descriptions of resources are achieved through the representation of the resource content in terms of concepts and structured data expressed by Web Ontology Language (OWL), and Resource Description Framework (RDF). The recent media interest around Wolfram Alpha (http://www.wolframalpha.com/), PowerSet (http://www.powerset.com) and Yahoo! SearchMonkey (http://developer.yahoo.com/searchmonkey/) show the expectations regarding the impact of this kind of Semantic Search.

On the other hand, we have also seen the successful adoption of ideas from IR to the problem of search in semantic (Web) data, which is due to the increasing size of the Semantic Web. Popular examples include the Linking Open Data (LOD) project (http://linkeddata.org/), the large body of data in forms of Microformats as well as RDFa data associated with text. Common to these scenarios is that the search is focused not on a document collection, but on semantic data, which may be possibly linked to or embedded in textual information.

Addressing complex information needs using textual documents and the large amount of semantic data on the Web today, are representing two main directions of Semantic Search that are featured in this special issue. Semantic Search is a broad concept studied from different perspectives by researchers in different communities. Firstly, this issue presents a general model, which considers these different perspectives and shows how semantics can be exploited throughout the search process for addressing both the tasks of document and data retrieval. Then, the content is divided into three parts: The first part contains research work, which specifically focuses on searching data (on the Web). As a special kind of data retrieval, the complex matching of service descriptions is also included in this part. The second part presents work that uses semantic data to improve the effectiveness of document retrieval. Finally, work considering the combined task of document and data retrieval, and relevance feedbacks between them is presented.

The first work covered by this issue is “SemSearchPro – Using semantics throughout the search process”. In order to collate various Semantic Search approaches and to better understand what the concept behind them entails, this work presents a general Semantic Search model, which is then extended to capture the notion of process-based Semantic Search, where semantics is exploited not only for processing the user queries, but may be involved in all steps of the search process – from constructing complex search queries to final result presentation. An implementation of this model is provided, based on which the practicability and usefulness of employing semantics throughout the search process are accessed through a task-based evaluation.
The data retrieval part of this issue starts with “Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine”, which presents a Semantic Web search engine supporting the search and browsing of Linked Data on the Web. The usage of RDF Web data – loosely also known as Linked Data – implies unique challenges such as scalability, unreliability, inconsistency and noise. Based on the six years working on this Semantic Web search engine project, the authors share their experiences and design choices made for SWSE. Since service descriptions are also available as semantic data, the search for services can be regarded as a special data retrieval task. The second paper contained in this part is “SAWSDL-iMatcher: A customizable and effective Semantic Web Service matchmaker”, which is about SAWSDL-iMatcher, a customizable Semantic Web service matchmaker. It supports a matchmaking mechanism that extends XQuery with various similarity joins for SAWSDL service discovery. Several matching strategies, including syntactic, semantic and statistical-model-based matching are employed to aggregate similarity values for various aspects of service descriptions including name, description text, and semantic annotation. The evaluation of this system shows that different matching strategies are suitable for different tasks and contexts, and that the effectiveness of SAWSDL service matching can be significantly improved by statistical-model-based matching.

In the document retrieval context, this issue firstly presents “Mapping queries to the Linking Open Data cloud: A case study using DBpedia” that uses semantic data to improve the understanding of Web search queries. This is demonstrated as a case study, which focuses on the use of semantic data contained in DBpedia, a major linking hub in the Linking Open Data cloud. The research problem studied here is how to map search engine queries to data in DBpedia to understand query keywords in terms of semantic concepts. A supervised machine learning-based method is proposed to determine which concepts in DBpedia are intended by the user query. This conceptual interpretation of the query may be used to provide contextual information, related concepts, or navigational suggestions to the user submitting the query. In the evaluation, the authors show that the proposed method significantly improves a baseline, which simply performs a lexical match between the queries and concepts. Next, an ontology-based approach for document retrieval is presented in “Semantically enhanced Information Retrieval: An ontology-based approach”, which uses semantic data contained in knowledge bases to enable semantically-enhanced search in large document repositories. Besides the ontology-based retrieval model, this work introduces a rank fusion technique that minimizes the undesired effects of knowledge sparseness. This is needed to deal with documents, for which no or only little semantic information is available. The last work along this line presented in “Semantic Web search based on ontological conjunctive queries” is called Serene, which considers the case of searching Web documents using queries more expressive than keywords, namely ontological conjunctive queries. For this, the system assumes the existence of ontologies, which adds ontological structure and semantics to Web pages. The authors allow for both attaching a meaning to Web search queries and Web pages, and for formulating and processing complex conjunctive queries over the Web. The actual search is realized by reduction to standard Web search on Web pages and logically completed ontological annotations such that standard search technologies can be leveraged as the main engine.

The last part of this special issue deals both with document and data retrieval. It starts with the study “Relevance Feedback Between Hypertext and Semantic Web Search: Frameworks and Evaluation”, which investigates the use of relevance feedback between document and data retrieval. While relevance feedback has shown to improve document retrieval performance, it has been used
over a single dataset. The idea here is to use relevance feedback from hypertext Web results to improve data retrieval on the Semantic Web, and results from the Semantic Web to improve the retrieval of hypertext Web data to create a “virtuous cycle” between these two tasks. The authors evaluate over a wide range of algorithms and options, and show how this virtuous cycle improves performance for deployed systems such as the Semantic Web search engine FALCON-S and Yahoo! Web search. Targeting information needs, which may require documents, structured data, or their combination as results, “Lightweight integration of IR and DB for scalable hybrid search with integrated ranking support” presents CE², a lightweight integration of IR and database technologies, which aims to address the problem of hybrid search. This search solution comprises a repository to store and index a large amount of semantic data as well as textual data in documents, a query engine that processes hybrid queries, and a flexible ranking scheme that is tightly integrated into the data and algorithms for query processing for reasons of performance.

A large amount of work on Semantic Search exists and this issue could only accommodate a small amount of selective articles. Through these, it presents a coherent overview of the main directions of Semantic Search, namely document retrieval, data retrieval and their convergence to hybrid search. Thus, Semantic Search as covered by this issue comprises a range of concepts and technologies, representing the convergence of Information Retrieval, database and Semantic Web technologies towards addressing complex hybrid information needs that may involve the retrieval of data, richly annotated documents and services, or other kinds of semantically-enhanced resource representations. Due to the large and increasing interests from research and industry in this topic, we expect this convergence to continue at a fast pace and research work to be focused on integrated solutions that can deal with the large amount of rich, heterogeneous as well as noisy information.

Rudi Studer, Institute AIFB, Karlsruhe Institute of Technology (KIT), Germany; studer@kit.edu

Yong Yu, Apex Data & Knowledge Management Lab, Shanghai Jiao Tong University, China, yyyu@apex.sjtu.edu.cn