DIESEL – Distributed Search over Large Enterprise Data

Ricardo Usbeck, Michael Röder, Muhammad Saleem, and Axel-Cyrille Ngonga Ngomo

AKSW Group, University of Leipzig, Germany
usbeck|roeder|saleem|ngonga@informatik.uni-leipzig.de

Project Goal. Data accessibility and integration belong to the main barriers to harnessing the full power of data in companies. In large companies, business-relevant data can be distributed across thousands of data silos in different formats. Existing semantic search solutions, e.g., Swoogle[1], mainly rely on extensions of text search (e.g., with domain-specific thesauri) and fail to make use of the semantics exposed in the data while interpreting queries. The DIESEL project[2] aims at developing a generic keyword search and question answering infrastructure for distributed and structured enterprise data. Our framework will exploit the distribution of the data to improve both the interpretation and the federated execution of user queries while considering the users' access rights. We will deliver an open-source version of the framework that implements search functionality as well as prototypical extensions of the partners' product suites and use case studies.

Project Methodology. DIESEL tackles the two drawbacks mentioned above in two ways. First, by developing a novel scalable approach to enable search through distributed Linked Data (LD). Instead of relying on enhanced text search, our approach will use the semantics of the user input to generate formal queries (i.e. SPARQL) out of keywords and natural language. A natural-language rendition of the interpretation of his query is represented to the user together with the search results. This enables the user to understand the way the system interprets his queries and to select an alternative interpretation.

Second, by Facilitating the deployment of the search solution over all enterprise data. DIESEL will deploy and adapt non-invasive scalable solutions for generating LD out of all types of data sources (i.e. unstructured, structured and semi-structured). *Our approach will enable companies to quickly integrate the LD paradigm into their information landscape without having to alter existing sources of information.*

Project Facts. The DIESEL project (Subject E!9367) has an overall runtime of 36 months which started in September 2015. It comprises 6 work packages with 18 deliverables. The consortium consists of 2 industry and 1 research partner. The Ontos AG, as coordinator, is a Swiss-based company founded in 2001 and is working in the area of web technologies, especially in the area of linked data, semantic web and natural language processing. The second partner is the

[1] http://swoogle.umbc.edu/

[2] http://diesel-project.eu/

metaphacts GmbH, a Germany-based company offering products, solutions and services for describing, interchanging and querying graph data, as well as a user-oriented open platform for visualizing and interacting with knowledge graphs. The research is mainly carried out by the University of Leipzig (ULEI) and its research group AKSW Group[3]. The research group focuses on research ranging from foundational research (Semantic Web, Link Discovery, Machine Learning, Data Integration, Knowledge Extraction) to applications that are deployed at industrial partners' sites. In the course of DIESEL, AKSW will focus on knowledge extraction from unstructured data as well as the integration of this data with existing structured data through linking. Moreover, ULEI will use its expertise on machine learning for structured data to detect patterns in integrated RDF data.

R & D activity. The research and development activities cover a broad range, starting from information extraction and conversion to RDF, federated SPARQL query processing to semantic search from any natural language input in multiple languages (DE, EN). Despite its very early stage, the DIESEL project partners were able to publish five articles, a.o. about efficient implementations for semantic similarity, federated querying via insights from predicates.[4] We strive for proofing the quality of the DIESEL by scientific publications as well as open source code and open benchmark data.[5]. Currently, the consortium starts implementing the defined architecture with respect to requirements elicited from user surveys, partner interviews and customer requirements.[6]

Session input and expectations. Within the EU networking session we focus on three points. First, we want to share our knowledge about state-of-the-art and up-to-date enterprise search engine behaviour and requirements. Thus, we will discuss with other participants their needs w.r.t. data sources, answering capabilities etc. This will support the development of the DIESEL search engine towards a versatile distributed search engine over large enterprise data. Moreover, we want to invite other researchers and practitioners to evaluate the usefulness of DIESEL towards possible follow-up projects in the context of H2020, Eurostars or national funding schemes. Second, we want to show early results (we will be in month 8) in terms of single component demonstrations, e.g., SPARQL federation and intuitive natural language interface, to receive early feedback as well as to point out future research directions. Finally, we want to identify complementary activities close to the DIESEL research and development area. For example, we are interested in crowd-based opportunities for information extraction support or novel techniques for searching and analysing large amounts of heterogeneous data.

[3] `http://aksw.org`

[4] `http://diesel-project.eu/results/publications/`

[5] `https://git.informatik.uni-leipzig.de/groups/diesel`

[6] `http://diesel-project.eu/results/deliverables/`