Requirements to Modern Semantic Search Engines

Ricardo Usbeck[1], Michael Röder[1], Peter Haase[2], Artem Kozlov[2], Muhammad Saleem[1], and Axel-Cyrille Ngonga Ngomo[1]

[1] AKSW Group, University of Leipzig, Germany
usbeck|roeder|ngonga@informatik.uni-leipzig.de
[2] metaphacts GmbH, Germany ph|ak@metaphacts.com

Abstract. Since the introduction of computing machines into companies and industries, searching large enterprise data is an open challenge including diverse and distributed datasets, missing alignment of vocabularies within divisions as well as data isolated in format silos. In this article, we report the requirements of commercial enterprises to the next generation of semantic search engine for large, distributed data. We describe our elicitation process to gather end user requirements, the challenges arising for real-world use cases as well as how such an implementation of this paradigm can be benchmarked. In the end, we present the design of the DIESEL search engine, which aims to implement the requirements of commercial enterprise to semantic search.

1 Introduction

Computing machines have been used for decades in companies and the amount of work that is supported by information technology is still growing. However, the problem of searching in large enterprise data is still an open problem due to grown data infrastructure of many companies. Several different software solutions are used for different tasks leading to a distributed and diverse landscape of data silos comprising different formats. The integration of all these data silos to enable enterprise wide search is a complex and time consuming task.

On the other hand, Linked Data technologies, built upon the Resource Description Framework (RDF), already showed their strength to extract, interlink and search over a variety of datasets. Still, the main disadvantage of RDF hindering its proliferation in the area of companies is its high entry barrier. Normal users need to be trained to cope with RDF data. Additionally, large parts of the existing software and database landscape would have to be adapted.

The main idea of the DIESEL project[3] is to use the strengths of Linked Data technologies to create an RDF-based layer on top of the diverse data silos. This layer enables the development of a search engine over all company and open data islands that makes use of the advantages of RDF without a) its drawbacks and b) without huge integration efforts. The consortium, which is composed of two

[3] http://www.diesel-project.eu

companies, a German and a Swiss-based company, as well as the Leipzig University, strives to provide an extensible open-source framework. In this article, we are going to present current requirements to modern semantic search engines. Our contributions in this article are as follows:

- First, we describe our use cases within the DIESEL project and to which extent these already point to novel research requirements.
- We present six up-to-date requirements analyses towards semantic-driven search over large, distributed enterprise as well as open data.
- We describe the influence of these requirements on our project and homogenize the raised issues to gain a concise overview and start outlining the DIESEL architecture.
- Consequently, we generate and collect matching benchmark definitions and datasets and present them to spark community contributions to the DIESEL platform.

To the best of our knowledge, we present the first systematic and industry-related as well as open data-based requirements specification for a large scale, semantic search engine.

2 Use Cases

In the following, we describe our three use cases within the DIESEL project.

Querying Enriched Encyclopedic Knowledge. DBpedia, YAGO as well as other open projects collect and systematize the world's knowledge and store it in a mostly structured way. However, accessing more complex knowledge from these sources remains a difficult task [19]. Thus, we will develop a generic search engine for encyclopedic knowledge that exceeds the capabilities of existing, text-based information retrieval approaches and truly understands the user intention while querying. This use case will make use of keywords and phrases and provide the user with state-of-the-art input comfort like auto-completion for query formulation. Furthermore, we will extend this knowledge by using unstructured streams (e.g. Twitter), tabular data (e.g. WHO) and other data sources which are not yet structured. Finally, we ensure that users can use any available knowledge base no matter of its location by implementing a federation layer—to query multiple RDF-based knowledge bases per input query—based on W3C standards like RDF and SPARQL 1.1 .

Wikidata. This use case focuses on how enterprises search can be supported by utilizing Wikidata. In contrast to the previous use case, the focus here is not limited to Wikidata itself, but rather its integration with enterprise data sources. Wikidata provides entity descriptions for approximately 17 million entities covering a variety of domains [22].[4] While wikidata is not an enterprise search data source per se, its potential for supporting and enriching enterprise search is immense. This is largely due to the fact that Wikidata develops more and more

[4] `https://www.wikidata.org/wiki/Special:Statistics`

towards a hub of identifiers for any kind of entities that enables the interlinking between disparate sources. In this context, the use case aims at enriching, contextualizing and integrating enterprise data with an open knowledge graph. Built on top of the Wikidata Knowledge Graph, the Wikidata Query Service[5] exposes this knowledge to the community and third-party developers through a scalable Web-based SPARQL endpoint, enabling queries such as "How did the population of Berlin develop over time", "Which countries are run by a female president", or "What are the most notable works displayed in the British Museum". The main focus will be, to allow customers to query the wealth of Wikidata in a way similiar to modern Web search engines without additional effort.

Medium-Large Enterprise Search and Knowledge Graph. Finally, we aim to leverage enterprise search by introducing a single point that enables querying all of a companies data as well as open data sources. This use case has as objective to ameliorate the search experience of employees in their everyday work. Here, we will extend the DIESEL engine to integrate the information from different sources where a semantically enriched federated search will be used as a single global search. Additionally, users will be able to visualise results in a structured and accessible graphical interface.

2.1 Elicitation Strategies

For our three use cases, we followed different elicitation strategies to account for the environmental situation of their future deployment.

Querying Enriched Encyclopedic Knowledge. First, we gathered academic requirements pertaining to search on encyclopedic data. Thus, we compiled the set of requirements behind the challenges Question Answering over Linked Data (QALD)[6] and Open Knowledge Base and Question-Answering (OKBQA)[7] by collecting:

- The motivation behind the benchmarks they provide. For example, to unify and extend existing as well as newly created datasets to compare the performance of systems.
- The current weaknesses of existing systems which took or are taking part in these challenges by analysing their internal structure and the corresponding publications.
- The drawbacks of existing datasets.

Overall, the elicitation led to the conclusion that current encyclopedic knowledge bases (YAGO, DBpedia) contain a large yet incomplete amount of valuable knowledge. However, a completion using knowledge gathered from data sources of another structure (in particular text and tables) would improve (1) the spectrum of queries that can be answered and (2) the completeness of the answers.

[5] https://query.wikidata.org/

[6] http://qald.sebastianwalter.org/

[7] http://www.okbqa.org/

Wikidata. Second, our German partner company used its community engagement in the Wikidata project, where it supported the development of the Wikidata Query Service based on the Blazegraph graph database.[8] As part of these activities, as well as through interviews with customers, which are anonymized, interested in using Wikidata and continued analysis of requirements expressed by the community (e.g. on the Wikidata mailing list), we elicited requirements for the Wikidata in enterprises use case.

Medium-Large Enterprise Search and Knowledge Graph. Finally, we analysed the requirements for enterprises which want to leverage knowledge graphs and search in their business environment. The collection of requirements for this use case was output of:

- Several informal interviews to the Swiss companies customers: this interviews had the objective to understand the organization's needs at management level, and to identify up to which extent they perceive the importance of integrating information within the company as well as with the German companies customers and leads.
- One formal interview with controlling personnel from a engineering contractor for rotating machines which seeks to manage and search its internal technical documentation: This interview focused on more technical approach, and had as objective to gather a single customer perspective on the information extraction aspects and the design of a user interface that could fulfil the needs.
- One formal interview with a customer of the German partner company, a large Swiss bank. The discussed potential use case involves analysts in a financial research department, who rely on search for their day-to-day analysis tasks, e.g., the task of predicting the development of inflation rates.
- A public online survey: This survey was designed with the objective of capturing most common requirements among different organisation, and also to understand how important it would be to have a global search among company data. The survey was created using Google forms, and is available at `https://t.co/8eBakzLLPK`[9]. It was distributed via our twitter account[10], through our LinkedIn contacts, and in the Search Engine Land LinkedIn group. We plan to leave the survey open considering that any future input can still be relevant for us.

2.2 Use Case-driven Requirements

Querying Enriched Encyclopedic Knowledge. The requirements for this use case focus on the data life cycle from extraction to distributed querying via SPARQL, quality of service and verbalization of answers. Table 1 lists the requirements from the QALD and OKBQA challenge in detail.

[8] `http://metaphacts.com/wikidata`

[9] At the 22nd April, we gathered 13 responses. In the online version you can also see preliminary reports.

[10] `https://twitter.com/project_diesel`

Table 1. Requirements for an Enhanced Encyclopaedic Search.

ID	Title	Description
1-1	Knowledge Extraction from unstructured sources	It must be possible to extract supplementary triples from text and use them to extend existing knowledge bases
1-2	Knowledge Extraction from semi-structured sources	It must be possible to extract supplementary triples from tables and XML documents and use them to extend existing knowledge bases
1-3	Federation of Queries	It must be possible to use knowledge from distributed RDF stores within one query within reasonable time
1-4	Runtime Efficiency	Search system for RDF data should be able to return answers within 3s.
1-5	Quality	Generated answers should have a high precision rather than a high recall to support trustworthy decision making
1-6	Verbalization of Answers	Users should see more than URIs.

Wikidata. The second use case focuses on the deployment of Wikidata in an corporate environment to enhance readily available data. Thus, the requirements evolve around interfacing between existing and future modules to enrich a company's workflow. Table 2 presents 9 requirements elicited from customer interviews.

Medium-Large Enterprise Search and Knowledge Graph. At the core of the DIESEL project is the design of a semantic search engine for large enterprise data. Thus, the last requirement collection aimed at covering as much aspects as possible in a concise way. For the sake of readability, we separated this use case requirement into four parts to account for the different elicitation methodologies. First, we present the results of the informal interviews with customers from our partner companies, see table 3.

Although there is a strong need for web-search, DIESEL will not focus on searching the Web since this is a different task than enterprise search. However, DIESEL will support searching web knowledge stored in internal data repositories which will have been transformed to RDF via our internal tools.

Second, we present the requirements from interviews with the rotary-machine producing companies controlling personal. Table 4 summaries the requirements. This customers also requires searching multiple documents, stored in various places and formats. Furthermore, this elicitation points out the importance of a context-aware knowledge extraction to reduce noise as well as the issue of corporate access control.

Third, we describe the requirements for a search system in the financial industry. Here, a Swiss bank details similar but also different aspects of enterprise search engines, see Table 5. These requirements introduce the issue of reusing existing thesauri and schemata, which is hard due to diverse matching prob-

Table 2. Requirements for using Wikidata within company search.

ID	Title	Description
2-1	Wikidata as an entity hub	Linking existing data with Wikidata identifiers to query existing data and knowledge graphs and link sofar isolated data sources.
2-2	Structured and unstructured data	Link together structured and unstructured data (e.g. Wikidata with the Wikipedia articles).
2-3	Elasticsearch	Due to the important role of entity search and the requirement to bridge with unstructured data, rich keyword search is essential. Of particular interest is support for Elasticsearch, as this is the search engine of choice of the Wikimedia projects.
2-4	Multilingualism	Use Wikidata as a useful resource for enabling multilingual search.
2-5	Contextualizing enterprise data	Linking enterprise data sources with Wikidata. Enable effective contextualization. Might require the combination of enterprise data with open data
2-6	Wikidata Ontology	Consider usage of Wikidata ontology while developing search interfaces.
2-7	Taxonomies	Besides formal ontologies, much of the knowledge is represented in hierarchical classification schemes and taxonomies. The use of these structures in search is essential
2-8	Hybrid searches	Search over Wikidata is very diverse, involving e.g. the following:- Entity searches- Structured queries ("What are the largest cities with a female mayor")- Property paths of unknown length (e.g. ancestor relations, territorial structures, taxonomic structures, part of relationships)- Discovery of links / paths between entities- Image Searches- Temporal and spatial data
2-9	Provenance	Management of trust and provenance is very important. Related aspects include the management of evidences, references, annotations etc.

lems. Moreover, financial research also relies on sentiment analysis of people, documents and market news. Thus an additional sentiment module would be advantageous. Finally, to enhance usability of the system it requires to represent results in a contextual and visual way to speed up understanding of the result.

At last, we introduce the results from our public survey which collected requirements from eight different companies. Table 6 presents the six derived requirements. Note here, that accessing already existing data in wiki-like systems, calendars and emails is at the core of user requirements. Again, we will not focus on search the Web, see above.

Table 3. Requirements for search in Knowledge Graphs of SMEs and larger enterprises via informal interviews with customers.

ID	Title	Description
3-1	Search for internal office documents and emails	Support DOC, PPT, XLS, PDF and MS[11] Exchange email Server.
3-2	Search in web	Support search in wiki, blogs and supplier web pages for guidelines, instructions and technical manuals.
3-3	Search by vocabulary	Support a common vocabulary with synonyms.
3-4	Datastore	Support MS SharePoint, file servers and the Web.
3-5	Search UI	Simple search slit using natural language and auto-completion.
3-6	Enhanced Search	Filter by vocabulary terms (concepts).
3-7	Filtering	Intuitive filtering by data type, time, owner, concepts.
3-8	Enrichment	Connect search results with other company (SAP ERP, CRM) and web data.
3-9	MashUp	Possibility to add mashup widgets, e.g., location-based information such as bus time table, events, point of interest
3-10	Trust	Possibility to specify only trusted sources.

3 System Design and Requirement Implications

In the following, we summarize and order the requirements to be able to derive a concise general architecture for our open source implementation. Moreover, we introduce possible existing frameworks and approaches to tackle the raised challenges. An overview of the DIESEL architecture is depicted in figure 1.

Knowledge Extraction for Enterprise Data First, a modern semantic search engine and respectively the DIESEL project needs to support the requested data sources. According to customers, each use case reports that DIESEL must be able to extract and search classical formats next to existing data sources:

- Support open data formats: XML, CSV, TSV, RDF, HTML.
- Support proprietary formats: MS Word, MS Excel.
- Support PDF which may require OCR technologies.
- Able to access existing data bases, mail servers, calendars, MS Share Point and media wiki.

Next to a semantic access to existing data sources, DIESEL needs to be able to truly understand the data at hand. Thus, the following criteria specify the capabilities of the data extraction modules:

- Extraction of domain-specific terminology (e.g. financial).

[11] MS stands for Microsoft

Table 4. Requirements for search in Knowledge Graphs of SMEs and larger enterprises via an interview with controlling personal from an engineering contractor for rotatary machines.

ID	Title	Description
3-11	Office documents	Support PDF documents.
3-12	None-text PDFs	Customer documents (old) are provided as scanned files in PDF format, thus, a OCR is required
3-13	MS Word documents	Customer documents can be provided in MS Word
3-14	MS Excel documents	Customer documents can be provided as MS Excel
3-15	Automatic meta-information extraction	Extract document name, location, metadata or document header and footer data
3-16	Search based on a taxonomy	A taxonomy of concepts with different labels is required.[12]
3-17	Multilingualism	Support documents in English and German.
3-18	Headers and footers	Since some of the project or document information is repeated in headers and footers, it creates critical noise that hinders to find detailed information fast.
3-19	Data source	Support WebDAV file system.
3-20	Data source	Support MS SharePoint.
3-21	Full text search	Resources should be searchable like Web search engines already do.
3-22	Faceted search	By using the taxonomy and metainformation, configurable facets should be provided.
3-23	Nearby terms	Support span searches within paragraphs.
3-24	Result preview	Enable search result snippets. (Answer verbalisation)
3-25	Result detail view	The user should have the possibility to click and open the document directly from the result view.
3-26	Access Control	Support Active Directory.

– Include metadata extraction (e.g. authors, creation dates) about the documents.
– Multilingual search capabilities per deployment for English and German.
– Extract entities such as places, organisation, persons, and also specific domain concepts (financial domain concepts, topics).

We will tackle these criteria by deploying data wrappers to RDF[13] such as Google Refine or RDF123 and will extend the systems where needed. Furthermore, we will make use of our own FOX-Framework [16], which is based on AGDISTIS [20], to extract RDF entities from unstructured data as well as deploying REX [2], a web-scale extraction system for heavily templated websites such as media wiki info boxes.

[12] Note, that the generation of the taxonomy takes a lot of time that nobody wants to invest.

[13] `https://www.w3.org/wiki/ConverterToRdf`

Table 5. Requirements gathered during an specification discussion from a Swiss bank.

ID	Title	Description
3-27	Search over unstructured data	Support PDF and HTML documents containing e.g., speeches about economic developments.
3-28	Concept and entity extraction	Effective search requires a precise identification of specific entities (such as places, organisation, persons) and concepts (financial domain concepts, topics).
3-29	Taxonomies and glossaries	Support existing glossaries, thesauri and taxonomies of relevant concepts and entities.
3-30	Sentiment analysis	The system requires to understand the sentiment of particular search terms.
3-31	Search over structured data	Support company internal sources with numerical data about economic indicators, such as inflation rates.
3-32	Contextualized result visualization	The result view should contain contextual information, e.g., how documents are related, their sentiment, covered topics etc. using, e.g., through graphs, heat maps, tag clouds.
3-33	Multilingualism	Support German and English.

Table 6. Requirements derived from public survey. Users entered these in a comment field.

ID	Title	Description
3-34	Separate security spaces	Account for the security level of the user.
3-35	Search on email	Support email search.
3-36	Search on media wiki	Support search on companies media wiki installation.
3-37	Calendars	Support search in various calendar formats.
3-38	Popular search engines	The first source of information in companies used is Web search engines.
3-39	Time	Reduce search time (less that 10 min for a global search).

DIESEL Search Engine Core After introducing a RDF layer on top of existing enterprise data, we need to develop a search algorithm that abides the following:

- The DIESEL search engine should provide high quality answers instead of a large number possibly irrelevant data.
- DIESEL should show the diverse interpretations of user input.
- The system has to reuse domain specific vocabularies and taxonomies.
- Multilingual queries: one DIESEL instance is required to process each language.
- The system should extend result with similar concepts.

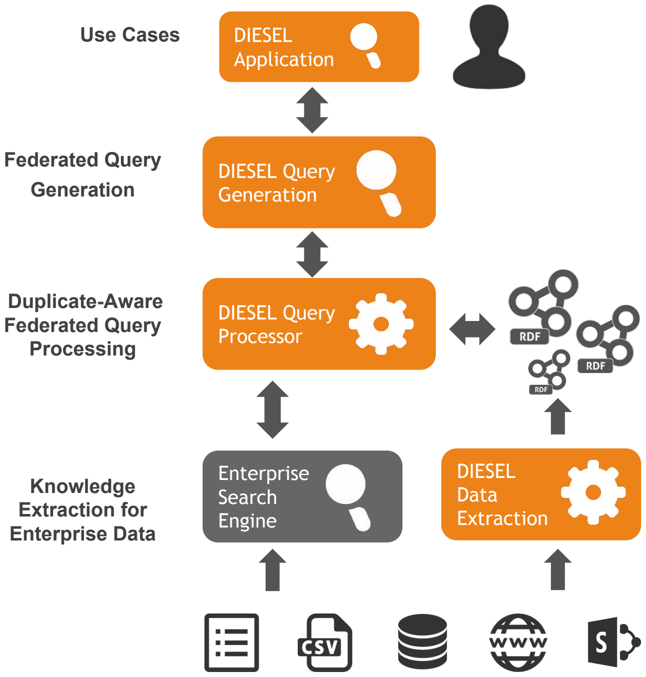

Fig. 1. Top-level Architecture of the DIESEL search engine.

- The system should provide distance between concepts provided in the search.
- The system should be able to rank results w.r.t. user intention.

As can be seen in many requirements, we will focus on a high precision instead of a high F-measure which would include also a high recall. In industrial environments a clear, expressive and trustworthy results seems to be more important than to no every possible correct answer. Our algorithm will be based on the underlying RDF data and leverage its graph-shape by using an algorithm similar to our own SESSA [8]. This algorithm is a high precision spread-activation algorithm over RDF graphs. Based on its high-quality result set, we will make use of existing Least-General-Generalization algorithms [6] to increase the answer coverage. Further, we will deploy a novel version of SPARQL2NL [10] to enable end-users to understand the internal representation of their input. Additionally, we will use make use of Ginseng[14], a modular semantic search engine to satisfy the need for similar concepts, proper result ranking and related concepts based on a similarity function. We will develop all DIESEL components with multilingualism in mind, with an initial focus on the languages required for the use cases, specifically German and English. Note, that the core DIESEL modules will not focus on query generation. This will be part of the use case implementations and the respective industrial exploitations.

[14] http://ginseng.aksw.org/

DIESEL Search over Distributed Data Although, we transformed existing data into RDF and our algorithm works on any RDF graph data, we need to be able to plug-in all data sources and query all endpoints at once. Thus, we need to implement a SPARQL federation layer, which satisfies the following user needs:

- Support querying different information silos within one query.
- Ensure access right restriction by enrolling a policy layer to the federation engine.
- Efficient execution.
- The system should be able to search in Solr.
- The system should be able to search in Elasticsearch.

We will extend the QUETSAL [12] federation engine which provides time efficient and effective source selection and ranking algorithms. QUETSAL already provides means to execute queries only on a subset of policy-restricted data sources via its SAFE extension [5]. The extension will consider time-efficient joins for federated queries and a duplicate aware federation. Finally, we will add capabilities to query state-of-the-art full-text search engines.

Use Case Specific Criteria Finally, there are requirements which only apply to certain use cases.

- Ability to include results from enterprise search engines.
- The search interface should provide a faceted search based in a taxonomy.
- The allows filtering by vocabulary terms (concepts).
- The system should also search into the Google Knowledge Graph

These extension will be built as extension to the previous two DIESEL modules, i.e., the query generation core and the query federation layer, to include other data sources, special filters or extend the Ginseng-based interface with novel widgets. Thus, DIESEL will remain modular and extensible.

Required Benchmarks To ensure a stable and guided development process, the development of a large-scale semantic search engine requires target key performance indicators (KPIs). Our survey yielded a number of measurements and values from which we used the maximal boundaries to push the limits of the DIESEL search engine further. Table 7 details the chosen modules and their respective KPIs.

4 Related Work

To develop a fully-fledged semantic search engine for large enterprise data tackles many fields. Here, we will focus on the field of keyword-based search as it is at the core of this survey.

Table 7. Target benchmark modules and performance indicators.

Benchmark	Method	Estimation/Measurement for Evaluation
Indexing	Collection of real-world log-files from DIESEL prototypes. Measuring the performance of single look-ups	<50ms per request
Similarity	Measure time-efficient implementation of similarity measures [9, 4]	<20ms per request
Auto-completion	Developing a benchmark for auto-completion.	Survey on user satisfaction
Query Expansion	Reuse and extension of benchmarks [14]	>0.65 F-measure
Ranking	Extend existing benchmarks to match DIESEL use cases [7].	>0.6 Accuracy
Query generation benchmark	Evaluation of query generation using QALD and other datasets [18]. Extraction, extension and evaluation of federated queries from [11]	>0.5 F-measure
Verbalization benchmark	Reuse and extend existing benchmarks from [10]	User study
Federation benchmark	Reuse benchmarks from [13].	<120 ms overall query runtime
Knowledge Extraction from Unstructured Data	Existing benchmarks based on the GERBIL platform [21]	>0.7 F-measure A2KB
Scalable Knowledge Extraction from Structured Data	Reuse existing benchmarks [2]	>0.9 Precision

Semplore [24] is the first known hybrid search engine by IBM. It combines existing information retrieval index structures and functions to index RDF data as well as textual data. Semplore focuses on scalable algorithms and is evaluated on an early Question Answering over Linked Data (QALD[15]) dataset.

Bhagdev et al. [1] describe an approach to hybrid search combining keyword searches, Semantic Web inferencing and querying. The proposed K-Search outperforms both keyword search and pure semantic search strategies. Additionally, a user study reveals the acceptance of the Hybrid Search paradigm by end users.

A personalized hybrid search implementing a hotel search service as use case is presented in [23]. By combining rule-based personal knowledge inference over subjective data, such as expensive locations, and reasoning, the personalized hybrid search has been proven to return a smaller amount of data thus resulting in more precise answers.

[15] `http://greententacle.techfak.uni-bielefeld.de/~cunger/qald/`

SINA [15] aims at answering a keyword question using different datasets. First, simultaneous disambiguation and segmentation is performed using Hidden Markov Models (HMM) and the Hyperlink-Induced Topic Search (HITS) algorithm. The resources found are used to construct an Incomplete Query Graph (IQG) constisting of disjoint sub-graphs. To build the federated SPARQL query that retrieves the results, the IQG's are connected using a Minimum Spanning Tree approach inspired by Prim's algorithm.

The work of Tran et al. [17] tackles the problem of keyword search over RDF data. More specifically, their work is concerned with mapping keywords to a list of ranked conjunctive queries, with a special focus on efficient inference of implied connections. To accomplish this, a top-k algorithm is proposed that computes the best query interpretations of the keyword query using bidirectional graph exploration. The interpretations are then scored and mapped to conjunctive queries.

Optique [3] aims at developing end-user oriented, semantic query interfaces over enterprise data sources by applying the techniques of OBDA (Ontology-based Data Access). With OBDA, end users can express queries over an ontology as conceptualization of the domain; mappings are employed to relate the ontology to the schemas of the underlying data sources. Optique provides an end-to-end OBDA platform, from end-user oriented query formulation to the actual query execution, but with a clear focus on relational data sources.

To the best of our knowledge, there is no open-soure semantic search engine satisfying a large subset of the requirements elicited earlier. Note, we are not focusing on exploring related commercial work.

5 Conclusion

In this article, we presented a concise requirements specification as well as benchmarking measures to develop an open data using, enterprise, semantic search engine over large data. We collected information from various sources, such as businesses, communities and academia. With these instruments at hand, we are going to develop the open-source DIESEL search engine together with our partners. Our intuition is that we cover enough use cases and domains to implement DIESEL in a way that will make it easy to generalize the project prototype to an industry-ready application.

In the future, we want to develop a first prototype and elaborate on the requirements to speed-up a close-to-reality implementation of our platform. We will integrate upcoming user requirements from interested parties in our development cycle and strive for an European dissemination of our efforts.

Acknowledgements. This work has been supported by Eurostars projects DIESEL (E!9367) and QAMEL (E!9725) as well as the European Union's H2020 research and innovation action HOBBIT (GA 688227).

References

1. R. Bhagdev, S. Chapman, F. Ciravegna, V. Lanfranchi, and D. Petrelli. Hybrid search: Effectively combining keywords and semantic searches. In S. Bechhofer, M. Hauswirth, J. Hoffmann, and M. Koubarakis, editors, *The Semantic Web: Research and Applications, 5th European Semantic Web Conference*, volume 5021 of *Lecture Notes in Computer Science*, pages 554–568. Springer, 2008.
2. L. Bühmann, R. Usbeck, A.-C. Ngonga Ngomo, M. Saleem, A. Both, V. Crescenzi, P. Merialdo, and D. Qiu. Web-Scale Extension of RDF Knowledge Bases from Templated Websites. In *The Semantic Web - ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, pages 66–81. Springer International Publishing, 2014.
3. M. Giese, A. Soylu, G. Vega-Gorgojo, A. Waaler, P. Haase, E. Jiménez-Ruiz, D. Lanti, M. Rezk, G. Xiao, L. Ö. Özgür, and R. Rosati. Optique: Zooming in on big data. *IEEE Computer*, 48(3):60–67, 2015.
4. J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 385–396, New York, NY, USA, 2014. ACM.
5. Y. Khan, M. Saleem, A. Iqbal, M. Mehdi, A. Hogan, P. Hasapis, A.-C. Ngonga Ngomo, S. Decker, and R. Sahay. SAFE: Policy aware SPARQL query federation over RDF Data Cubes. In *Semantic Web Applications and Tools for Life Sciences(SWAT4LS)*, 2014.
6. J. Lehmann and L. Bühmann. Autosparql: Let users query your knowledge base. In *Proceedings of ESWC 2011*, 2011.
7. V. Lopez, A. Nikolov, M. Fernández, M. Sabou, V. S. Uren, and E. Motta. Merging and ranking answers in the semantic web: The wisdom of crowds. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *ASWC*, volume 5926 of *Lecture Notes in Computer Science*, pages 135–152. Springer, 2009.
8. D. Lukovnikov and A.-C. Ngonga-Ngomo. Sessa - keyword-based entity search through coloured spreading activation. In *NLIWoD@ISWC*, 2014.
9. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 1–8. ACM, 2011.
10. A.-C. Ngonga Ngomo, L. Bühmann, C. Unger, J. Lehmann, and D. Gerber. SPARQL2NL - Verbalizing SPARQL queries. In *Proc. of WWW 2013 Demos*, pages 329–332, 2013.
11. A. Nikolov, A. Schwarte, and C. Hätter. Fedsearch: Efficiently combining structured queries and full-text search in a sparql federation. In H. Alani, L. Kagal, A. Fokoue, P. Groth, C. Biemann, J. Parreira, L. Aroyo, N. Noy, C. Welty, and K. Janowicz, editors, *The Semantic Web ISWC 2013*, Lecture Notes in Computer Science, pages 427–443. Springer Berlin Heidelberg, 2013.
12. M. Saleem, M. I. Ali, R. Verborgh, and A.-C. Ngonga Ngomo. Federated query processing over linked data. In *Tutorial at ISWC*, 2015.
13. M. Saleem, A.-C. Ngonga Ngomo, J. X. Parreira, H. Deus, and M. Hauswirth. Daw: Duplicate-aware federated query processing over the web of data. In *Proceedings of ISWC2013*, 2013.
14. S. Shekarpour, K. Höffner, J. Lehmann, and S. Auer. Keyword query expansion on linked data using linguistic and semantic features. In *7th IEEE International Conference on Semantic Computing, September 16-18, 2013, Irvine, California, USA*, 2013.

15. S. Shekarpour, A.-C. Ngonga Ngomo, and S. Auer. Question answering on interlinked data. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1145–1156. International World Wide Web Conferences Steering Committee, 2013.
16. R. Speck and A. N. Ngomo. Ensemble learning for named entity recognition. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference*, pages 519–534, 2014.
17. T. Tran, H. Wang, S. Rudolph, and P. Cimiano. Top-k exploration of query candidates for efficient keyword search on graph-shaped (rdf) data. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 405–416. IEEE, 2009.
18. C. Unger, C. Forascu, V. Lopez, A. N. Ngomo, E. Cabrio, P. Cimiano, and S. Walter. Question answering over linked data (QALD-5). In *CLEF*, 2015.
19. R. Usbeck. Combining Linked Data and Statistical Information Retrieval. In *11th ESWC, PhD Symposium*, 2014.
20. R. Usbeck, A. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. AGDISTIS - graph-based disambiguation of named entities using linked data. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference*, pages 457–471, 2014.
21. R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *24th WWW conference*, 2015.
22. D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, Sept. 2014.
23. D. Yoo. Hybrid query processing for personalized information retrieval on the semantic web. *Knowledge Base Systems*, 27:211–218, 2012.
24. L. Zhang, Q. Liu, J. Zhang, H. Wang, Y. Pan, and Y. Yu. Semplore: An IR Approach to Scalable Hybrid Query of Semantic Web Data. In *ISWC*, pages 652–665, 2007.