SCAIView – A Semantic Search Engine for Biomedical Research Utilizing a Microservice Architecture

Jens Dörpinghaus[1,2], Jürgen Klein[1], Johannes Darms[1], Sumit Madan[1], and Marc Jacobs[1]

[1] Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany
[2] jens.doerpinghaus@scai.fraunhofer.de

Abstract. Biological and medical researchers explore the mechanisms of living organisms and tend to gain a better understanding of underlying fundamental biological processes of life. To tackle such complex tasks they constantly need to gather and accumulate new knowledge by performing experiments and studying scientific literature. We will present the novel semantic search engine "SCAIView" for knowledge discovery and retrieval and, additionally, discuss the most recent paradigm shifts in communication technologies, which leads to a completely new architecture that improves scalability, achieves better interoperability, and also increases fault-tolerance.

1 Introduction

Biological and medical researchers are interested in exploring the mechanisms of living organisms and gaining a better understanding of underlying fundamental biological processes of life. To tackle such complex tasks they constantly gather and accumulate new knowledge by performing experiments, and also studying scientific literature that includes results of further experiments performed by researchers. Existing solutions are mainly based on the methods of biomedical text mining to extract key information from unstructured biomedical text (such as publications, patents, and electronic health records).

Especially in the field of biomedical sciences, we have a long history of developing applications that solve the above mentioned tasks. For instance, SCAIView[3] is an information retrieval system that allows semantic searches in large textual collections by combining free text searches with the ontological representations of automatic recognized biological entities (see Hodapp et al. [5]). SCAIView was used in many recent research projects, for example regarding neurodegenerative diseases [4] or brain imaging features [6]. Furthermore, it was also used for document classification and clustering [3]. Another important

[3] https://www.scaiview.com/ (an academia version is freely available at http://academia.scaiview.com/academia/)

real-world task is the creation of biological knowledge graphs that is tackled by the BELIEF environment [9]. It assists researchers during the curation process by providing relationships extracted by automatic text mining solutions and represented in a human-readable form [10]. At the core of both technologies several implementations of the methods of biomedical text mining are in place.

In this poster we will present the recent development of SCAIView, and how SCAIView (as well as BELIEF) evolved using the same core technologies to an interoperable software system.

2 SCAIView architecture

To keep up with the state-of-the-art technologies and to be prepared for integration of novel and game-changing developments, we migrated the SCAIView ecosystem from a large monolith to microservice-based system. It allows us to reuse parts for different purposes and the data itself can be easily processed, shared and accessed. Additionally, the new system also allows us to focus on FAIR (Findable, Accessible, Interoperable, and Reusable) principles, introduced in [11], that are becoming a standard in the biological scientific community.

The microservice infrastructure of SCAIView is an ecosystem of three main services: Core, API, and Indexer (Figure 1), which communicate through the message broker (Apache ActiveMQ). The core fulfills various important tasks to persist, retrieve, and process data. Beside further text mining microservices, there are also specialized microservices such as BEL Commons Professional, which allows to validate text-mined biological entities and relationships, that are shared by BELIEF and SCAIView ecosystems. SCAIView's user interface itself is a web-based microservice application running on Apache Tomcat communicating via REST-API calls with the backend. The visualization of the document corpus includes document elements that are stored and represented as semantic digital assets (SDA) (Jacobs et al. [7]). The SDA represent various semantically-enriched domain models that can be binary data like images or plain-text such as natural language. The corpus itself is pre-processed and stored in a document store.

The *Document Store* is based on Apache Accumulo and Apache Solr. The first one is used to persist raw results of the text mining pipelines. This allows us to compare and validate the development of old and new text mining components really fast, which is necessary in the research area. The latter one contains SDAs such as the document text, recognized semantic concepts, and further metadata that is needed for fast retrieval. SCAIView can also handle multiple text mining and knowledge discovery pipelines by communicating through the message broker. Common steps are the usage of a DocumentDecomposer, Lemmatizer, JProMiner for named entity recognition. Other text processing components, such as UIMA Ruta-based components (see [8]) or ChemoCR (see [12]) can be used on demand and be easily integrated into processing pipelines.

Search queries and knowledge discovery in SCAIView is linked to ontology and terminology data. Semantic searches are a combination of free text search and entities represented in ontologies or terminologies. For instance, SCAIView

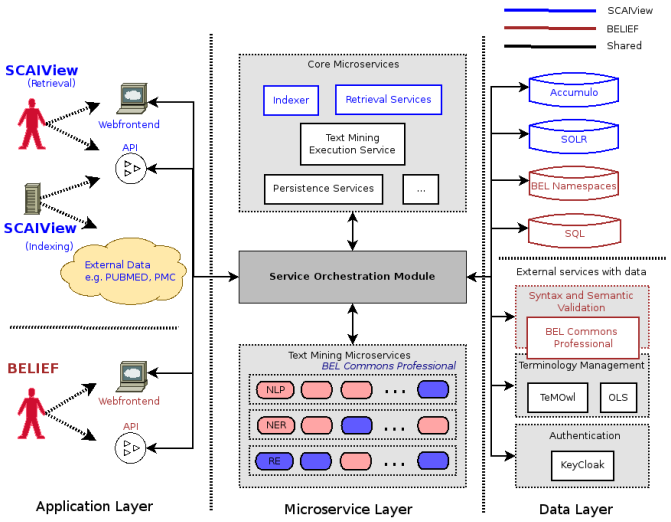

Fig. 1. The shared architecture for the semantic search engine (SCAIView) and the semi-automatic knowledge graph creation environment (BELIEF). It consists of three different layers: application, microservice, and data layer. The BEL network-related microservices are called *BEL Commons Professional*.

includes Alzheimer's Disease Ontology (ADO), BioMarker terminology, drug names, the Hypothesis Finder and many more. These resources are displayed in a tree format and can be used to make detailed, faceted search queries and to perform statistical analysis on the retrieved document corpus. The access to these resources is provided by our internal-hosted OLS service (*Ontology Lookup Service* [2]) and the upcoming TeMOwl (Terminology Management based on OWL) service.

In general, SCAIView is developed to handle any kind of document corpus but currently we focus on the biomedical research area. Therefore, as input we use databases such as PubMed 2017 [1] that contains around 27 million abstracts and PMC 2017[4] that includes around 2 million biomedical-related full-text articles. Following [7] and [5] the processing of huge data is not only possible, but also very efficient and the microservice infrastructure is highly scalable.

3 Conclusion

Although several risks and problems have to be faced, we are sure that positive advantages of implementation of a microservice system do outweigh. For both applications, SCAIView as well as BELIEF, several microservices are used and shared for purpose of data retrieval, data persistence, and text mining. The latter are *classical* microservices, whereas the retrieval and persistence services are more general microservices. Additionally, the microservices in the data layer can also be traditional webservices such as the terminology management or authentication systems. We benefit from a highly scalable and fault-tolerant environment for data processing. Furthermore, the system is flexible enough to easily add or remove microservices from the processing pipeline. The continuous

[4] https://www.ncbi.nlm.nih.gov/pmc/

delivery process for externally-developed software like OLS or Keycloak is not an issue anymore. An additional benefit is the safe and fast switching from one technology to another: TeMOWl and OLS can be used at the same time for multiple instances of SCAIView.

References

1. Coordinators, N.R.: Database resources of the national center for biotechnology information. Nucleic acids research **45**(Database issue), D12 (2017)
2. Côté, R.G., Jones, P., Martens, L., Apweiler, R., Hermjakob, H.: The ontology lookup service: more data and better tools for controlled vocabulary queries. Nucleic acids research **36**(suppl_2), W372–W376 (2008)
3. Dörpinghaus, J., Schaaf, S., Fluck, J., Jacobs, M.: Document clustering using a graph covering with pseudostable sets. In: Computer Science and Information Systems (FedCSIS), 2017 Federated Conference on. pp. 329–338. IEEE (2017)
4. Emon, M.A.E.K., Karki, R., Younesi, E., Hofmann-Apitius, M., et al.: Using drugs as molecular probes: A computational chemical biology approach in neurodegenerative diseases. Journal of Alzheimer's Disease **56**(2), 677–686 (2017)
5. Hodapp, S., Madan, S., Fluck, J., Zimmermann, M.: Integration of UIMA Text Mining Components into an Event-based Asynchronous Microservice Architecture. In: Proceedings of the LREC 2016 Workshop "Cross-Platform Text Mining and Natural Language Processing Interoperability". pp. 19–23. European Language Resources Association (ELRA), Portorož, Slovenia (2016)
6. Iyappan, A., Younesi, E., Redolfi, A., Vrooman, H., Khanna, S., Frisoni, G.B., Hofmann-Apitius, M.: Neuroimaging feature terminology: A controlled terminology for the annotation of brain imaging features. Journal of Alzheimer's Disease **59**(4), 1153–1169 (2017)
7. Jacobs, M., Hodapp, S., Dörpinghaus, J.: SDA: Towards a novel Knowledge Discovery Model for Information Systems. In: Proceedings of the 11th IADIS International Conference Information Systems 2018. pp. 300–302. IADIS (2018)
8. Kluegl, P., Toepfer, M., Beck, P.D., Fette, G., Puppe, F.: Uima ruta: Rapid development of rule-based information extraction applications. Natural Language Engineering **22**(1), 1–40 (2016)
9. Madan, S., Hodapp, S., Senger, P., Ansari, S., Szostak, J., Hoeng, J., Peitsch, M., Fluck, J.: The BEL information extraction workflow (BELIEF): evaluation in the BioCreative V BEL and IAT track. Database **2016**, baw136 (oct 2016). https://doi.org/10.1093/database/baw136, http://database.oxfordjournals.org/lookup/doi/10.1093/database/baw136
10. Szostak, J., Ansari, S., Madan, S., Fluck, J., Talikka, M., Iskandar, A., De León, H., Hofmann-Apitius, M., Peitsch, M.C., Hoeng, J.: Construction of biological networks from unstructured information based on a semi-automated curation workflow. Database : the journal of biological databases and curation **2015** (2015). https://doi.org/10.1093/database/bav057
11. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data **3** (2016)
12. Zimmermann, M.: Chemical structure reconstruction with chemocr. In: TREC (2011)