Fraunhofer
SCAI

FRAUNHOFER INSTITUTE FOR ALGORITHMS AND SCIENTIFIC COMPUTING SCAI

UIMA-HPC

HOME | WORKFLOWS | ME

Fraunhofer Institute for Algorithms and Scientific Computing SCAI
Schloss Birlinghoven 1
53757 Sankt Augustin, Germany

Contact:
Dr. Marc Jacobs
Phone +49 2241 14-4013
marc.jacobs@scai.fraunhofer.de

www.scai.fraunhofer.de

Our UIMA-HPC project collaboration partners:

JÜLICH FORSCHUNGSZENTRUM

scapos

TAROS Chemicals

UIMA-HPC is a collaborative project funded by the German Federal Ministry for Education and Research (BMBF), Förderkennzeichen: 01IH11012A.

FINDING AND DELIVERING KEY INFORMATION HIDDEN IN PHARMACO-CHEMICAL PATENTS

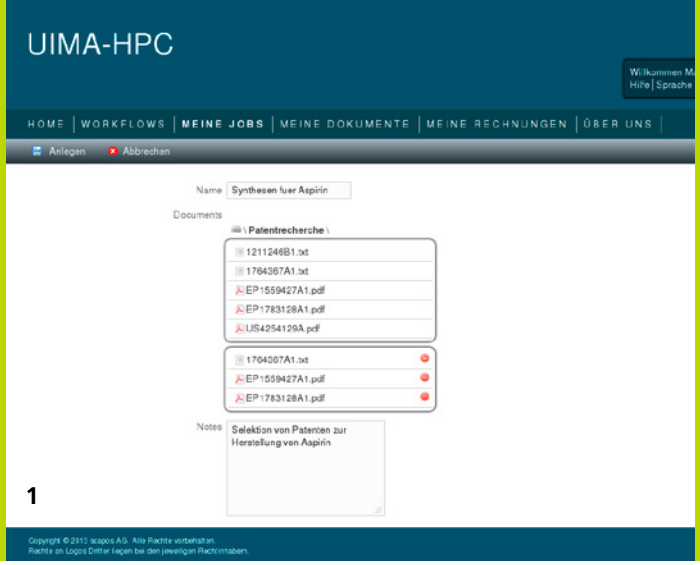
Finding the knowledge needle in the data haystack

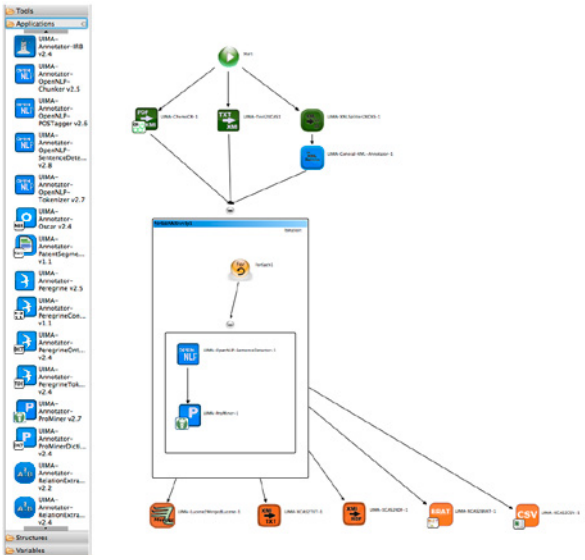
98% of human knowledge is in the form of digital records. Most of these records are unstructured – hence extracting information is a time and cost consuming task. UIMA-HPC combines two key technologies that make it possible to get best time-to-solution results: Unstructured Information Management Architecture (UIMA) and Uniform Interface to Computing Resources (UNICORE). The first provides all UIMA-HPC workflows with software-component-based parallelization, the latter facilitating access to high performance computing (HPC) systems and the utilization of hardware-level parallelism. Combined for the first time as part of this project, these technologies synergize for maximum efficiency in solving information extraction tasks on huge unstructured document collections.

Pharmaco-chemical patents – a particular challenge

The development of new chemicals or pharmaceuticals is critically dependent on a prior in-depth analysis of the published patents in this field. This is a cost- and time-consuming step when done by a human reader. One specific goal of the research project UIMA-HPC is to automate and hence speed-up the process of knowledge mining in pharmaco-chemical patents.

The UIMA-HPC project is developing a service system that will enable the rapid analysis of patents from the Pharmaceutical and Bio-chemical sectors, exploiting sector-tuned mining engines and domain expertise. UIMA-HPC provides fast and massive parallel processing of patent records in multiple formats, including Optical Character Recognition of PDFs and the





detection of chemical structure information from drawings within patents.

Reading all common file types

Our workflows allow the combination of input from different source file types such as PDF, TXT, CSV, XML, DOC and MySQL databases that are all transformed into the UIMA internal data object CAS. Thus, all file types can be processed by identical analysis engines.

Analyses of the data

UIMA-HPC has developed a system of Pipelets – each with a specific application and use-case. All Pipelets are compatible with one another and once combined, create a full workflow. Existing Pipelets represent traditional Natural Language Processing tools such as Sentence Detection, POS-Tagging and Chunking complemented by various systems for (chemical) Named Entity Recognition (NER). We also provide special Zoning Pipelets that can easily be configured and which extract certain sections from documents such as abstracts, summaries or – more complex – paragraphs where a synthesis description is found.

UIMA-HPC Web Portal

Our service portal provides secure access to a personal workspace as part of the UIMA-HPC system. Pre-defined complex workflows appear as simple Black-Box applications, annotated with a description of their function, and convenient use is enabled via a drag&drop interface. Submitted workflows indicate their status (pending, running, done) and output files are automatically stored online in secure UNICORE storages that are only accessible by the corresponding client. Our asynchronous scale-out allows users to concentrate on other tasks while documents are processed – no online monitoring is necessary.

Output formats

UIMA-HPC supports many different output formats that allow a seamless integration into existing knowledgebases. All identified entities can be stored as simple CSV files, which can be imported into common spreadsheet applications. Following RDF standards we also write n3 files that can be uploaded into an existing triplestore. For full-text analysis we provide a Lucene-based writer for storing user's data. Different formats can easily be integrated into existing workflows.

1 *Service Portal providing easy-to-use document management for automated orchestration of analysis workflows comprising UIMA processes, realised as Grid-Beans, executed on remote HPC systems via Unicore.*