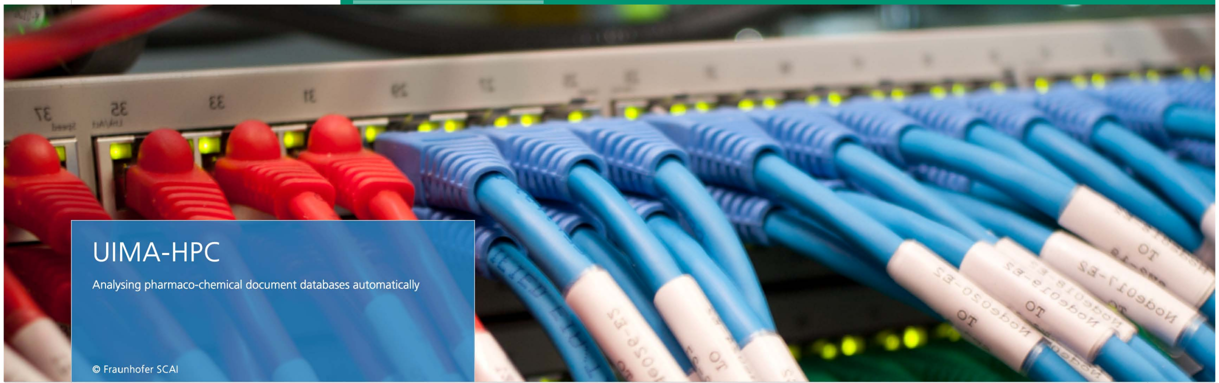


UIMA-HPC

Analysing pharmaco-chemical document databases automatically

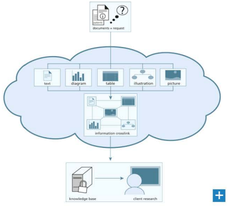
© Fraunhofer SCAI

About UIMA-HPC

Finding the Knowledge Needles in the Data Haystack

The explosion of unstructured data available for research and development is a general phenomenon, but it has already become a performance defining factor in the medical and Biotechnology / Pharmaceutical areas: without ICT-based support tools for automated mining of document databases, determination and retrieval of strategically important scientific and business information is either untenable or becomes a significant drain on manpower resources. The situation in Pharmaceutical and bio-chemistry sectors is made more extreme by the reliance on multi-modal information in publications and documents as chemical structures are not just represented in text form but also as structure diagrams.

A particular, representative focal point is patent search in the pharmaco-chemical context: mining of patent documents requires a combination of text mining based on domain-specific vocabularies and ontologies combined with information extraction from (printed versions of) chemical structure diagrams. With databases containing millions of complex documents, the automated data analysis process is one whose computational requirements require high-performance computing and in order to meet the needs of the many industrial small and medium enterprises in the sector, a solution delivery approached based on remote service computing as offered by Cloud and SaaS solutions.



© Fraunhofer SCAI

Diagram visualizing the UIMA-HPC workflow.

Analysing pharmaco-chemical document databases automatically

The UIMA-HPC project aims to realize an HPC-based solution for the automated analysis of multi-modal pharmaco-chemical document databases, taking the patent-search use-case as an initial solution design driver. The combination of text and structure analysis is an innovative approach, but will be based on an existing and well-tested data analysis architecture: the Unstructured Information Management Architecture (UIMA). UIMA is a software architecture which specifies component interfaces, design patterns and development roles for creating, describing, discovering, composing and deploying multi-modal analysis capabilities. The UIMA specification is being developed by a technical committee at OASIS.

The UIMA-HPC approach centres on the workflows for the automated annotation of a document corpus, the workflow comprising analysis components within the UIMA architecture. The individual »annotation engines«, such as text-mining of a document or analysis of diagrams within a document based on Optical character recognition (OCR), are of a computational complexity such that parallelization at the level of the heterogeneous »node« of a modern HPC system is highly appropriate, meaning parallelization for deployment on multi-core and/or GPU-accelerated processors. Handling the large quantity of documents – and the related load-balancing issues created by the diversity of computational complexity relating to individual documents – to be analyzed by independent instantiations of the annotation engines for the workflow is handled at the level of the nodes of the HPC compute system as a whole and will be realized within an adaptation of the Unicore software system.

Note:

Apache UIMA, UIMA are trademarks of The Apache Software Foundation.

Contact

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Funded project

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

Partners

The Consortium is led by Fraunhofer SCAI and includes

- Forschungszentrum Jülich
- scapos AG
- Taros Chemicals GmbH

Partners

- Fraunhofer SCAI

Forschungszentrum Jülich GmbH

Taros Chemicals GmbH & Co. KG

scapos AG

Fraunhofer SCAI

Fraunhofer SCAI is internationally one of the leading groups in the challenging field of information extraction from the biomedical literature. SCAI has a special focus on biological name recognition and disambiguation of synonyms.

Fraunhofer SCAI brings its expertise in information extraction, and the already developed software tools to the project. The SCAI approach is based on methods from Computational Linguistics and Bioinformatics. The established name recognition of biomedical and chemical terms is of central importance for the extraction of textual information. The assignment of different synonyms for defined entities such as genes or chemical compounds (disambiguation) and the recognition of speech varieties are highly relevant for the pharmaceutical industry. The Institute SCAI has developed in cooperation with Aventis Pharma an internationally competitive platform for the identification and normalization of named entities (ProMiner). In addition to the extraction from texts, a novel prototype for the reconstruction of chemical structures from images has been developed in the recent years (chemoCR). Chemical depictions can be found in many scientific publications and in chemistry related patents. The pre-processing and analysis of the layout of complex documents, such as scientific papers and patents, is of utmost importance. SCAI currently has a unique selling proposition with the combination of the tools ProMiner (text) and chemoCR (pictures). In the field of text mining, SCAI could show the excellent quality of the solution in public competitions. Worldwide only few academic groups are conducting research on the problem of chemical image mining.

http://www.scai.fraunhofer.de/en.html

Related Links:

- Forschungszentrum Jülich
- scapos AG
- Taros Chemicals GmbH & Co. KG

Forschungszentrum Jülich GmbH

The Jülich Supercomputing Centre (JSC) provides computation time on supercomputers, information technology tools and knowhow for the Research Centre Jülich and all over Europe. The widely varied knowledge about software development for and in operation of HPC systems characterizes the national high-performance research centre JSC.

For UIMA-HPC the JSC provides the operation of UNICORE services and computing time. The Grid middleware UNICORE provides the intuitive and efficient usage of Grid resources by using e.g. use-case specific workflows.

http://www.fz-juelich.de/ias/jsc/EN

Related Links:

- Fraunhofer SCAI
- scapos AG
- Taros Chemicals GmbH & Co. KG

Taros Chemicals GmbH & Co. KG

Taros Chemicals is an established provider of chemical research located in Dortmund, Germany. Using 13 years of experience successfully working for major pharmaceutical, biotechnical, agricultural, chemical and personal-care companies has left an impact.Taros Chemicals works in a seamless partnership with its clients, covering services traditionally provided by in-house divisions, such as drug and process research and chemical custom synthesis. We can thereby considerably lower client's process costs by shortening the R&D phase for drugs and processes enabling our biotech customers to reach their goals without building up/expanding their own chemistry division preventing backlogs in the synthetic process. This in turn allows clients to pursue their core business issues creating increased opportunities for discovery and development.

www.taros.de

Related Links:

Fraunhofer SCAI
scapos AG
Forschungszentrum Jülich

scapos AG

In order to strengthen marketing and sales of their products, the Fraunhofer-Institute for Algorithms and Scientific Computing SCAI initiated the launch of scapos AG; the Fraunhofer Society is one of the scapos shareholders. The company also offers its services to other Fraunhofer institutes and research organizations.

The scapos product portfolio is characterized by

- innovative mathematical algorithms
- development in close cooperation with industrial customers
- optimized performance on advanced hardware architectures
- proven cost and time reduction for customers

The scapos office is located at the campus of Fraunhofer's institutional center, Schloss Birlinghoven, which allows us to maintain close connections to scientists and product developers.

http://www.scapos.com

Related Links:

Fraunhofer SCAI
Taros Chemicals GmbH & Co. KG
Forschungszentrum Jülich

Publications

2012 2011

Paper von der eChallenges e-2012 Conference 17-19 Oktober, Lissabon,

UIMA-HPC – Application Support and Speed-up of Data Extraction Workflows through UNICORE

Publications

2012 2011

Artikel aus dem inSiDE Journal (Vol. 9 No. 2 • Autumn 2011) des Gauss Centre for Supercomputing (GCS)

UIMA-HPC: High-Performance Knowledge Mining

7th German Conference on Chemoinformatics, 2011 Goslar (FHG)

Download PDF, 1.1 MB

CGW11 - the Eleventh Cracow Grid Workshop, 2011 Cracow (FZJ)

Download PDF, 1.3 MB

UIMA-HPC

Analysing pharmaco-chemical document databases automatically

Events

Event 2013

- International Supercomputing Conference ISC'13
  16-20 Juni Leipzig
- UNICORE Summit
  18 Juni Leipzig
- 3. HPC-Status-Konferenz Gauß-Allianz
  5-6 September Dresden

Events 2012

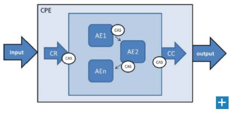
- 244th ACS National Meeting & Exposition
  19-23 Dezember Philadelphia, USA
- UNICORE Summit
  30-31 Mai Dresden
- ChemAxon's 8th European User Group Meeting
  22-23 Mai Budapest, Ungarn

Share SHARE TWEET SHARE SHARE PRINT

Follow us

Technical

Methods



© Fraunhofer SCAI

Fig. The input is converted into CAS by a collection reader (CR), further processed by a number of analysis engines (AE), and finally written back by a collection consumer (CC).

The distinctive feature of UIMA-HPC is the flexible generic approach which makes it applicable to any kind of UIMA-Pipelines and workflows thereof as well as any kind of compute resources, which are available.

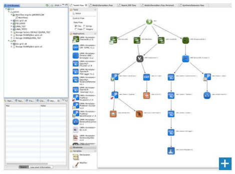
UIMA pipelines are the basic building blocks of information extraction workflows. Apache UIMA provides a native Java framework for mining unstructured data. An UIMA application is organized as a Collection Processing Engine (CPE) that consists of an UIMA Collection Reader (CR), one or more UIMA Analysis Engines (AEs) and one Collection Consumer (CC). The analyzed artifact (e.g. text or binary data) is stored in the internal UIMA data structure Common Analysis Structure (CAS). The framework architecture also provides convenience methods for serializing CAS objects (XCAS) to store them persistently on hard disk. These stored XCAS files can then again be read by a CR. In our implementation we exploit this procedure to transport data between physically separated hardware nodes.

Information extraction from chemical patents

The goal of the research project UIMA-HPC is to automate and hence speed-up the process of knowledge mining in patents. Multi-threaded analysis engines, developed according to UIMA (Unstructured Information Management Architecture) standards, process texts and images in thousands of documents in parallel. UNICORE (UNiform Interface to COmputing Resources) workflow control and execution features capabilities make it possible to dynamically allocate resources for every given task to gain best cpu-time/real-time ratios in an HPC environment.

All UIMA components (CPE, CR, AE and CC) are specified via XML file format descriptors, which contain consistent predefined internal routes. For a Grid system we need a dynamic handling of network paths. Therefore we use the UIMAFit implementation to generate all XML specifications at run-time of an UIMA pipeline. The necessary import of uniform resource identifiers (URIs) in all Java classes of UIMA can be dynamically adapted to any location using UIMAFit. All our integrated pipelines are provided as a Java archive files (jar) and run platform independent on different operating systems. The framework architecture UIMA makes it possible to easily integrate existing software and also replace AEs within different UIMA pipelines.

Existing Components



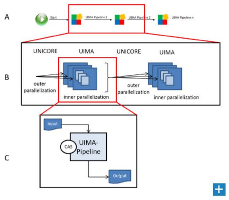
© Fraunhofer SCAI

A workflow that demonstrates all UIMA-Components available at this time. Collection Reader Pipelets are shown in green, Analysis Engines in blue and Consumers in orange, respectively. UIMA-View converter is shown in dark-grey. Open Source software such as OpenNLP components are seamlessly integrated into one workflow together with proprietary software components (ProMiner, chemoCR) by sharing the same UIMA-TypeSystem.

Examples of implemented UIMA pipelines to process documents with medical and chemical content

INPUT	INTEGRATED 3RD PARTY SOFTWARE	FUNCTION	ANNOTATIONS	OUTPUT
PDF	CLI abbyy finereader	OCR	SourceDocument Information	XCAS
PDF	PDFbox, iText	Text extraction	SourceDocument Information	XCAS
XCAS	ProMiner	Dictionary based Annotation	Chemistry, Diseases, Genes	XCAS
XCAS	Linda	Machine Learning (ML) based Annotation	Diseases, Genes, IUPAC-terms	XCAS
XCAS	OSCAR	Dictionary and ML based annotation of chemical terms	Chemical terms	XCAS
XCAS	iText, PDFBox	Generating annotated PDF		Enriched PDF

UIMA and UNICORE



© Fraunhofer SCAI

Fig. Complete architecture of the coupling between UIMA and UNICORE

In order to make UIMA pipelines available on distributed heterogeneous resources to be accessible through UNICORE they have to meet certain requirements:

- Installed on the target system,
- Executable as stand-alone applications,
- No hard-coded paths in file descriptors.

The overall architecture is shown in Figure 2. As UIMA is a native Java library it is cross platform compatible and can be installed on UNIX and Microsoft Windows based servers. The prerequisite is an installation of Java Virtual Machine 6 or higher.

A UIMA pipeline is provided as a Java jar archive, which has to be available on a server's file system. The Input and output data format is defined in XML, it is called serialised CAS objects (C). This must be unified to be free in the choice of annotations and their order in a workflow. The Java archive is made available through UNICORE by defining it as an application resource (B). Upon execution the jar archive is called by UNICORE via a system call using the standard arguments of the Java virtual machine. The XML application configuration files support any number of arguments that can be defined prior to execution separately for every job on the client side. UIMA provides multithreading of embedded components. This allows to exploit all cores of a node in the provides multithreading of embedded components. This allows to exploit all cores of a node in the execution environment.

Share

SHARE TWEET SHARE SHARE

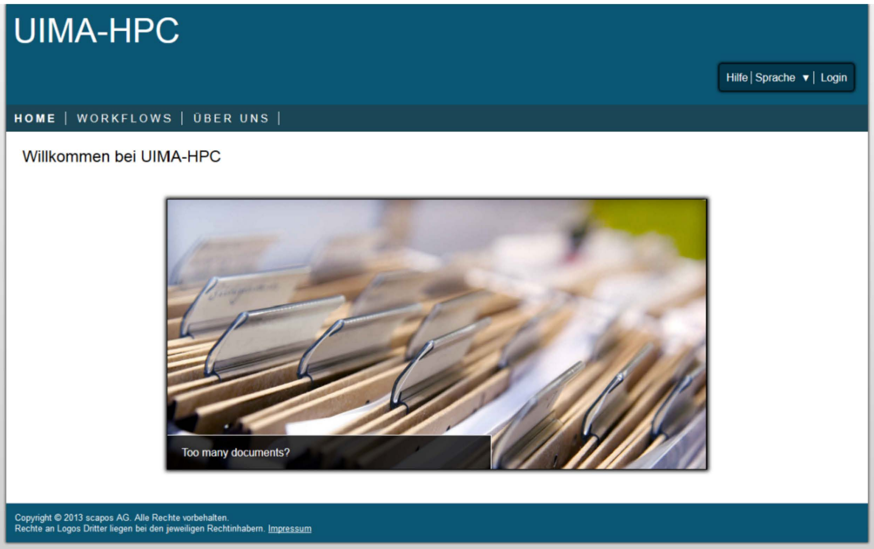
PRINT

Follow us

Demo

If your are interested in a free demo access, please contact us via E-Mail.

WebPortal



Willkommen bei UIMA-HPC

Too many documents?

Copyright © 2013 scapos AG. Alle Rechte vorbehalten.
Rechte an Logos Dritter liegen bei den jeweiligen Rechtinhabern. Impressum

Homepage

© Fraunhofer SCAI

●○○