FROM BIG DATA TO SMART KNOWLEDGE – TEXT AND DATA MINING IN SCIENCE AND ECONOMY

COLOGNE, FEBRUARY 23 TO 24 2015

PROGRAM

CONTENT

3 Welcome letter
4 Program overview
5 Conference dinner
8 Abstracts
17 General information
18 Local organising committee
19 Organising partners
20 Map

SPONSORING

The conference is supported by

averbis
text analytics

TEMIS

WELCOME LETTER

Dear Ladies and Gentlemen,
'Big data' is a widely-used term that refers to a major trend in the world of technology. There are a number of questions that providers of scientific information, i.e. libraries need to address in order to keep pace with this trend:

- How can we assist and support people in using scientific information beyond simply providing access to PDF files?
- How can unstructured data (text and multimedia content) be used in ways that go beyond the 'traditional' acts of reading papers and watching videos?
- What technologies are available to accelerate the processes of data and information gathering, information aggregation and knowledge discovery?
- How can future-oriented library services support research in the fields of biomedicine, engineering, technology and business?

All these points will be discussed during the next two days here at the Hyatt Hotel in Cologne. With a carefully chosen selection of presentations from leading academic experts and industrial researchers, the conference will present the current situation and examine how this may develop in the future, illustrating how we can make the shift from big data to smart knowledge.

Thank you for being part of the international conference "From Big Data to Smart Knowledge – Text and Data Mining in Science and Economy" – we wish you inspiring presentations and discussions!

On behalf of the organising institutions,

Ulrich Korwitz
ZB MED – Leibniz-Information Centre for Life Sciences

Martin Hofmann-Apitius
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)

MONDAY, FEBRUARY 23

11.00 **Registration – Coffee and Snacks**

13.00 **Conference Opening and Welcome Notes**
Ulrich Korwitz
ZB MED – Leibniz Information Centre for Life Sciences, Cologne/Bonn (Germany)

13.15 **Keynote**
Datastewardship for Discovery
Prof. Barend Mons
Erasmus Medical Centre, University of Rotterdam and Department of Human Genetics, Leiden University Medical Centre (the Netherlands)

14.00 **Session 1 – Applied Research**

Text mining as a key to information: an interdisciplinary perspective on the exploration of patents and historical textbooks
Prof. Christa Womser-Hacker
Department of Information Science & Natural Language Processing, University of Hildesheim (Germany)

Sentiment Analysis and Opinion Mining in Product Reviews: Fine-grained Analysis and Cross-Linguality
Dr. Roman Klinger
Visiting Professor for Theoretical Computational Linguistics, Institute for Natural Language Processing, University of Stuttgart (Germany)

15.15 **Coffee Break & Networking**

PROGRAM OVERVIEW

16.00 **Session 2 – Application Examples**

Text and data mining @Roche: an industry perspective
Dr. Markus Bundschus
Head Scientific & Business Information Services, Roche Diagnostics GmbH, Penzberg (Germany)

Access to knowledge: Text mining and information extraction in the German National Library
Reinhard Altenhöner
German National Library, Frankfurt/Main

Modelling hypothetical knowledge: Capturing and representing scientific speculation in text
Prof. Martin Hofmann-Apitius
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), St. Augustin (Germany)

19.00 **Conference Dinner**

February 23, 19.00 h

Conference Dinner at the Sion Brewery

Sion BRAUHAUS SEIT·1318

The Sion Brewery
Located in the historic center of Cologne, you can enjoy draft beer – the inevitable "Kölsch" is brewed only in the region of cologne – as well as good and solid food. You can choose between different specialities like the "Halve Hahn" (rye bread roll with Dutch cheese and mustard), the "Kölsche Kaviar" (blood sausage) or "Rievkooche" (potato cake).

TUESDAY, FEBRUARY 24

9.00 **Opening Day 2**
Prof. Martin Hofmann-Apitius

9.15 **Keynote**
Resolving phenotypes to standard representations: a complex task
Dr. Dietrich Rebholz-Schuhmann
University of Zuerich (Switzerland)

10.00 **Session 3 – Fundamental Research**

Visual mining - interpreting image data
Prof. Stefan Rüger
Knowledge Media Institute, The Open University, Milton Keynes (UK)

10.45 **Coffee Break & Networking**

11.15 **Session 4 – Translational Aspects**

Pragmatic text mining: From literature to electronic health records
Prof. Lars Juhl Jensen
NNF Center for Protein Research, University of Copenhagen (Denmark)

Extraction from scientific text of causal and correlative relationships used in systems biomedicine models of disease
Dr. Juliane Fluck
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI),
St. Augustin (Germany)

12.45 **Lunch & Networking**

PROGRAM OVERVIEW

14.00 **Session 5 – Enabling Technologies**

Large-Scale Patent Classification at the European Patent Office
Dr. Philipp Daumke
Averbis AG, Freiburg (Germany)

Text Mining and Compliance - Supporting access to complex regulatory legislation by natural language processing
Stefan Geissler
TEMIS Germany, Heidelberg (Germany)
Matthias Leybold
Deloitte Consulting AG Switzerland

Impact of developments in big data analytics for new use cases
Dr. Anton Heijs
datasciencesets, Gouda (the Netherlands)

15.30 **Closing Remarks**
Prof. Martin Hofmann-Apitius

15.45 **End with Coffee and Farewell Snacks**

Keynote, day 1

Datastewardship for Discovery

Prof. Barend Mons
Erasmus Medical Centre, University of Rotterdam and Department of Human Genetics, Leiden University Medical Centre (the Netherlands)

Knowledge Discovery across data resources is hampered by the lack of standards and the poor adoption of existing standards by stakeholders. Data Interoperability overcomes the barriers of syntactic access with semantic use in one implementation. Optimal Interoperability is only attained when access and use can be completely automated: programming and interfaces conform to standards that specify consistent syntax and formats; and data are associated with metadata and terminology identifiers and codes that support computational aggregation and comparison of information that resides in separate resources.

Many European research programmes and public private partnerships already make significant investments in data infrastructure to make data better Findable, Accessible, Interoperable and ultimately Reusable (FAIR), but without coordination such as provided by ELIXIR the large numbers of stakeholders and programmes within Europe will drive fragmentation and overlapping investments in data management, stewardship, analytics and technology approaches. Through the implementation of community adopted and ELIXIR endorsed standards and, importantly, a European wide framework of experts and a credible supporting organisation, ELIXIR will drive the coordination efforts both at national and international level. ELIXIR is an Open Infrastructure - it will not "own" or ‚control' the data resources in Europe but provide a coordinated Backbone that enables and assists partners (e.g. other ESFRI Research Infrastructures) to make use of existing solutions and connect and interoperate their resources. Sustained infrastructure services for e.g. identifier management, data access and mappings between resources drive "standards as the community driven default" and enable long-term data management according to the FAIR principles (data should be Findable, Accessible, Interoperable and Reusable).

ABSTRACTS

Session 1 – Applied Research

Text mining as a key to information: an interdisciplinary perspective on the exploration of patents and historical textbooks

Prof. Christa Womser-Hacker
Department of Information Science & Natural Language Processing, University of Hildesheim (Germany)

The elicitation of meaningfull information from huge text and data collections has gained increasing interest in research and economy. Several disciplines participate in this extensive challenge. From the information science perspective it is important to understand information needs in detail, know about working environments of the target user groups, and mainly their understanding of information that should be extracted and visualized by systems in order to support users' research goals and strategies. Text Mining techniques powered by Natural Language Processing, Statistics, and Machine Learning must be adapted to the special requirements underlying the tasks of the target audience represented by different domain experts. In my talk, I will report on ongoing interdisciplinary projects (funded by the Leibniz Association). Even though both domains seem to be very different: Trend Mining in the patent domain (T4P project) and mining for visions and interpretations of the world as portrayed in historical textbooks for children („Welt der Kinder"), both projects are conducted in close collaboration between Digital Humanists, Patent experts, Computer and Information Scientists, and Computational Linguists.

Sentiment Analysis and Opinion Mining in Product Reviews: Fine-grained Analysis and Cross-Linguality

Dr. Roman Klinger
Visiting Professor for Theoretical Computational Linguistics, Institute for Natural Language Processing, University of Stuttgart (Germany)

Sentiment Analysis and Opinion Mining is often phrased as a text classification task or, in a more fine-grained setting, as text segmentation to extract specific phrases denoting aspects under discussion and evaluating phrases with polarities assigned by an author.
In that sense, sentiment analysis is, from a methodological point of view, similar to other information extraction tasks like named entity recognition or relation extraction. However, the specification of text segments to be detected is different and hard to be stated.

In this talk, I give a short introduction to different challenges and applications of coarse-grained and fine-grained sentiment analysis and different methods to address those. I will then introduce a model for joint detection of evaluating phrases and associated aspects as mentioned in product reviews as they are available from shop websites like Amazon. I conclude with a short overview on our recent work in training models across different languages.

ABSTRACTS

Session 2 – Application Examples

Text and data mining @Roche: an industry perspective

Dr. Markus Bundschus
Head Scientific & Business Information Services, Roche Diagnostics GmbH, Penzberg (Germany)

Even though Text and Data Mining is part of the technology portfolio for many years in the industry, only recently it is shifting from being a niche player towards becoming an integral part of business critical processes. The range of applications is huge and diverse in pharmaceutical companies – from traditional use cases such as drug and biomarker discovery, analyzing clinical trials, or optimizing biotechnological production processes, to finding key opinion leaders, among others. Given the unprecedented growth of scientific knowledge represented in written documents, there are currently not so many alternatives in the future to that automated processing technique.
In this talk we discuss our strategy how to successfully implement text mining projects in a challenging industrial setting. We outline important design criteria that have to be critically selected to ensure broad use of these powerful technologies. We highlight selected use cases and discuss open research questions that would be important to be tackled from the industry perspective.

Access to knowledge: Text mining and information extraction in the German National Library

Reinhard Altenhöner
German National Library, Frankfurt/Main

The German National Library (DNB) is faced with a massive increase of born-digital publications in their collections. In order to offer access to these materials for users the library evaluates ways to use automated data analysis processes. Here, the library also has to consider established access systems and indexing rules and routines. At the same time the metadata infrastructure becomes more and more global and data analysis and data linkage is becoming increasingly important and potentially valuable to reuse and enrich existing classification information. Some methods have been taken productive, others are still in the project phase or simply experimental.

The presentation provides an overview of the approaches DNB follows so far and highlights potentials, which may have a future impact for the further development of the library and information infrastructure.

Modelling hypothetical knowledge: Capturing and representing scientific speculation in text

Prof. Martin Hofmann-Apitius
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), St. Augustin (Germany)

Modelling hypothetical knowledge: Capturing and representing scientific speculation in text Speculative statements communicating experimental findings are frequently found in scientific articles, and their purpose is to provide an impetus for further investigations into the given topic. Automated recognition of speculative statements in scientific text has gained interest in recent years as systematic analysis of such statements could transform speculative thoughts into testable hypotheses. We describe here a pattern matching approach for the detection of speculative statements in scientific text that uses a dictionary of speculative patterns to classify sentences as hypothetical. To demonstrate the practical utility of our approach, we applied it to the domain of Alzheimer's disease and showed that our automated approach captures a wide spectrum of scientific speculations on Alzheimer's disease. Subsequent exploration of derived hypothetical knowledge leads to generation of a coherent overview on emerging knowledge niches, and can thus provide added value to ongoing research activities.

ABSTRACTS

Keynote, day 2

Resolving phenotypes to standard representations: a complex task

Dr. Dietrich Rebholz-Schuhmann
University of Zuerich (Switzerland)

Capturing single phentype traits or the full phenotype description is a complex task due to the large number of traits that form the phenotype, and due to the different types of qualities linked to individual phenotypes (e.g., lack of an organ, insufficient function, increase/decrease of a physiological parameter). For human, mouse and other model organisms, specific resources have been produced (e.g., Human phenotype ontology, mouse phenotype ontology) to capture the description of a phenotype. This talk will give an overview on the use of public resources to denote a phenotype, on solutions to use model organism data in combination and on the limitations of linking genes to diseases through by data integration using terminologies and ontologies.

Session 3 – Fundamental Research

Visual mining - interpreting image data

Prof. Stefan Rüger
Knowledge Media Institute, The Open University, Milton Keynes (UK)

Like text mining, visual media mining tries to make sense of the world - albeit by analysing pixels instead of words. This talk looks at some important aspects of visual media mining, for example, near-duplicate detection, multimedia indexing and the role of machine learning in typical mining tasks such as analysing food images. Near duplicate detection allows us to link the real world with databases based on a snapshot from a smart-phone or a wearable camera. Multimedia indexing creates a structure from media repositories for retrieval and visual search engines. The talk will end by looking into the crystal ball to explore what might be learned from automatically analysing tens of thousands of hours of TV footage.

Session 4 – Translational Aspects

Pragmatic text mining: From literature to electronic health records

Prof. Lars Juhl Jensen
NNF Center for Protein Research, University of Copenhagen (Denmark)

Text mining is rapidly becoming an essential tool for biomedical data mining. The literature is a vast source of knowledge, most of which is not captured by existing structure databases. Electronic health records (EHRs) are another underused textual data source, the mining of which has the potential for revealing unknown disease correlations and for improving post-approval monitoring of drugs. In my presentation I will introduce a pragmatic approach to mining the biomedical literature for drugs, proteins, subcellular compartments, tissues, diseases, and associations among them. I will also describe how we apply the same techniques to identify adverse reactions of drugs from the clinical narrative in electronic health records.

Extraction from scientific text of causal and correlative relationships used in systems biomedicine models of disease

Dr. Juliane Fluck
Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), St. Augustin (Germany)

In order to build networks for systems biology from the literature, an UIMA based extraction workflow using various named entity recognition processes and different relation extraction methods has been composed. The Unstructured Information Management architecture (UIMA) is a Java-based framework that allows assembling complicated workflows from a set of NLP components. The new system is processing scientific articles and is writing the open-access biological expression language (BEL) as output. BEL is a machine and human readable language with defined knowledge statements that can be used for knowledge representation, causal reasoning, and hypothesis generation. BEL-based disease models can now be generated by automatically processing hundreds of thousands of full text documents, effectively speeding up model generation in systems biomedicine at unprecedented scale.

ABSTRACTS

Session 5 – Enabling Technologies

Large-Scale Patent Classification at the European Patent Office

Dr. Philipp Daumke
Averbis AG, Freiburg (Germany)

In the era of Smart Data and the explosion of data volume of all kind, organizations seek for leveraging such data - being it patent information, research literature social media data etc. - for competitive advantage and to help achieving their strategic aims. The process of search, filtering and categorization of large data sets go typically far beyond simple keyword search. Semantic technologies paired with machine learning approaches from artifical intelligence are a promising approach to support more fine-granular analysis of data.
The European Patent Office and Averbis recently went into collaboration for the pre-classification of incoming patent applications (use case 1) and re-classification of existing classification schemes (use case 2). In this cooperation, various services are provided with the aim of automatically assigning patent applications to the right departments and automatically allocating existing patents with new CPC codes. The solution is based on complex linguistic and semantic analyses, as well as statistically-based machine learning processes. Up to 250.000 incoming patents shall be classified per year and categorized in up to 1.500 categories. In this talk, we want to present both use cases together with some technical background about the applied language technologies.

Text Mining and Compliance - Supporting access to complex regulatory legislation by natural language processing

Stefan Geissler
TEMIS Germany, Heidelberg (Germany)
Matthias Leybold
Director Analytics, Deloitte Consulting AG Switzerland

Legal regulatory frameworks are often of a complexity that makes it hard even for the expert reader to digest the necessary information. Verifying and ensuring compliance with the relevant legislation then often means having to handle countless mutually dependent regulations that refer to one another as well as to specific sophisticated technical terminologies. Natural language processing can provide essential support in managing and accessing information of this type. The presentation describes a use case around the FATCA („Foreign Account Tax Compliance Act") legislation that has been successfully applied in the banking industry. The take home message may be that certain legal frameworks are today of a complexity that benefits from or even requires modern natural language processing technologies to make it accessible to the intended audience.

Impact of developments in big data analytics for new use cases

Dr. Anton Heijs
datasciencesets, Gouda (the Netherlands)

Developments in big data analytics technology and machine learning techniques have made data and text-mining more powerful. This enables new use cases in pharma/biotech and healthcare where large amounts of structured and unstructured data can be used. Combining analysis of structured data (table or image data) with text data can enable faster and richer insights. The algorithmic and technological developments of big data analytics enabling scalable processing and analysis with an overview of all the data will be discussed. The value of drill down approaches especially using visualization will be presented. Also the importance of detecting trends, patterns and semantics of especially large text data sets is analyzed. Some new use case that benefit from scalable processing with an overview of all data will be discussed including the requirements and created value coming from such use cases. Especially in the medical and life sciences domain these development will have a huge impact although there are still many challenging complexities that need to be addressed in the near future.

Conference venue

Hyatt Regency Cologne
Kennedy-Ufer 2a
50679 Cologne
Tel.: +49 221 828 1234
The conference will take place in Rheinsaal 2 on the mezzanine floor.

Conference office

Opening hours:
Monday, 23 Feb, 10.00-18.00 h
Tuesday, 24 Feb, 8.30-17.00 h
Location: In front of the Rheinsaal 2, Hyatt Hotel, mezzanine floor
E-mail: marketing-team@zbmed.de
Tel.: +49 221 8281 1542

WIFI

You will receive a free wifi access with your registration at the conference office.

Twitter-Hashtag

#tdm15

LOCAL ORGANISING COMMITTEE

Chair and program:
Prof. Martin Hofmann-Apitius (martin.hofmann-apitius@scai.fraunhofer.de)
Ulrich Korwitz (korwitz@zbmed.de)

Project management and marketing:
Ulrike Ostrzinski (ostrzinski@zbmed.de)

Press contact and website:
Juliane Tiedt (tiedt@zbmed.de)

Finances and billing:
Jürgen Gärtner (gaertner@zbmed.de)

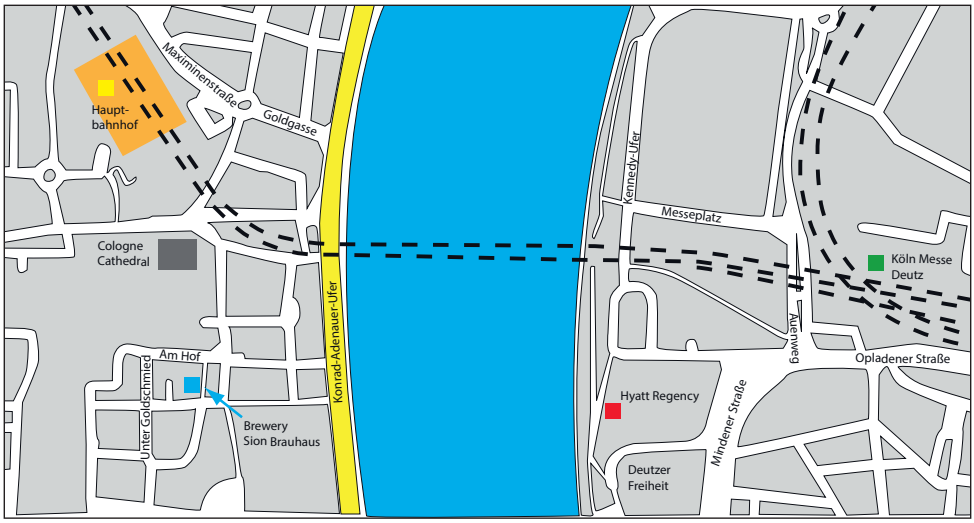
ORGANISING PARTNERS

Goportis – Leibniz Library Network for Research Information
Goportis is the name of the Leibniz Library Network for Research Information which consists of the three German National Libraries: ZB MED – Leibniz Information Centre for Life Sciences (Cologne/Bonn), the TIB (National Library of Science and Technology, Hanover), and the ZBW (National Library of Economics – Leibniz Information Centre for Economics, Kiel/Hamburg). The three partners work together to support scientific working processes by providing research-based services, for example by opening up access to research data, developing virtual research environments and semantic applications, and offering various other types of assistance.
www.goportis.de

GOPORTIS
Leibniz Library Network
for Research Information

Fraunhofer Institute for Algorithms and Scientific Computing SCAI
The Fraunhofer Institute for Algorithms and Scientific Computing SCAI conducts research in the field of computer simulations for product and process development, and is a prominent corporate partner in the industrial and science sectors. SCAI designs and optimizes industrial applications, implements custom solutions for production and logistics, and offers calculations on high-performance computers. Our services are based on industrial engineering, combined with state-of-the-art methods from applied mathematics and information technology.
www.scai.fraunhofer.de

Fraunhofer
SCAI



Conference Venue:

Hyatt Regency Cologne
Kennedy-Ufer 2a
50679 Cologne
Phone: +49 221 828 1234

Conference Dinner Location:

Brewery „Sion Brauhaus“
Unter Taschenmacher 5-7
50667 Köln
Phone: +49 221 2 57 85 40

Railway Station

Cologne Main Station (Köln Hauptbahnhof)

Railway Station

Cologne Messe/Deutz (Bahnhof Köln Messe/Deutz)

#tdm15