„FROM BIG DATA TO SMART KNOWLEDGE – TEXT AND DATA MINING IN SCIENCE AND ECONOMY“ (KÖLN, 23.–24. FEBRUAR 2015)

von Bruno Bauer

Am 23. und 24. Februar 2015 fand im Hyatt Hotel in Köln die internationale Konferenz *„From big data to smart knowledge – text and data mining in science and economy“* statt, die von zirka 60 Expertinnen und Experten aus Bibliotheken, Universitäten und weiteren Forschungseinrichtungen sowie der Industrie besucht wurde. Veranstaltet wurde die Tagung vom Goportis – Leibniz-Bibliotheksverbund Forschungsinformation gemeinsam mit dem Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI. Im Goportis – Leibniz-Bibliotheksverbund Forschungsinformation kooperieren die drei deutschen zentralen Fachbibliotheken TIB (Technische Informationsbibliothek, Hannover), ZB MED (ZB MED – Leibniz-Informationszentrum Lebenswissenschaften, Köln/Bonn) und ZBW (Deutsche Zentralbibliothek für Wirtschaftswissenschaften – Leibniz-Informationszentrum Wirtschaft, Kiel/Hamburg).

Big Data, Smart Knowledge, Text Mining, Data Mining sind Begriffe, die aus Wissenschaft und Forschung nicht mehr wegzudenken sind. Durch die zunehmende Digitalisierung und die digitale Verfügbarkeit von Daten gewinnen Text und Data Mining an Bedeutung und verändern die Wissenschaft. In seiner Eröffnungsrede wies **Ulrich Korwitz** (ZB MED – Leibniz-Informationszentrum Lebenswissenschaften, Köln/Bonn) darauf hin, dass das Thema mittlerweile auch Gegenstand von politischen Stra-

tegiepapieren der Forschungspolitik ist; zu nennen sind etwa die *G8 Open Data Charter* (https://www.gov.uk/government/publications/open-data-charter/g8-open-data-charter-and-technical-annex) oder *The Hague Declaration on Knowledge Discovery in the Digital Age* (http://thehaguedeclaration.com/wp-content/uploads/sites/2/2015/04/Liber_DeclarationA4_2015.pdf), in der postuliert wird, dass „*the right to read*" auch „*the right to mine*" inkludieren muss.

Die Keynote des ersten Tages wurde von **Barend Mons** (Universität Rotterdam und Universität Leiden) zum Thema *„Datastewardship for Discovery"* gegeben. Er wies auf die zunehmende Bedeutung von Daten für die Forschung hin, zeigte aber auch auf, dass mangels Standardisierung bei Syntax, Formaten und Metadaten die Nachnutzung von Daten oft nur erschwert möglich ist. Optimal wäre eine maschinenlesbare Veröffentlichung von Daten analog zu Zeitschriftenartikeln, am besten mit einer Art von Impact Factor. Während Daten bisher zumeist nur im Kontext von Texten in Zeitschriftenartikeln veröffentlich werden, wünscht sich Mons in Zukunft den freien Zugang zu Daten – Stichwort Open Access. Ermöglicht werden soll diese Entwicklung durch die Umsetzung der *FAIR Guiding Principles* der *JDDCP* (*Joint Declaration of Data Citation Principles*). *FAIR* steht in diesem Zusammenhang für *findable*, *accessible*, *interoperable*, *re-usable*, und bietet Empfehlungen zu persistenten Identifikatoren, zur Nutzung von einheitlichen Protokollen, zu Workflows und Formaten sowie zu Metadaten. Um Daten gemäß diesen Richtlinien bereitstellen zu können sind der Erwerb entsprechender Qualifikationen in der wissenschaftlichen Ausbildung sowie eine ausreichende Finanzierung für *Data Stewardship* Voraussetzung. Optimal wäre es, fünf Prozent der für ein Projekt zur Verfügung stehenden Mittel für *Data Stewardship* zu nutzen. Weiters wäre es wichtig, für Forschende Anreize zu bieten, ihre Forschungsdaten zu publizieren und mit anderen zu teilen.

Anschließend sprach **Klaus Tochtermann** (Deutsche Zentralbibliothek für Wirtschaftswissenschaften) zum Thema *„On the Evolution of Semantic Technologies in Scientific Libraries"*. Bei Recherchen erweist es sich oft als Problem, dass ohne den Einsatz von semantischen Technologien bei synonymen Begriffen keine vollständigen Sucherergebnisse erzielt werden können. Zur Lösung dieses Problems können *Semantic Graphs* beitragen. Dabei werden auch Ergebnisse geliefert, in denen eine große Ähnlichkeit zwischen Dokumenten und Knowledge Base besteht, obwohl der Suchbegriff selbst in diesen Dokumenten nicht aufscheint. Treffermengen können auch durch die Erstellung von Konkordanzen zwischen unterschiedlichen Thesauri vergrößert werden, wobei hier die Gefahr besteht, dass auch viele nicht relevante Treffer angezeigt werden.

Roman Klinger (Universität Stuttgart) sprach über „*Sentiment Analysis and Opinion Mining in Product Reviews: Fine-grained Analysis and Cross-Linguality*“. Am Beispiel einer Uhr wurde dargestellt, dass es sehr individuelle Gründe gibt, warum man etwas mag oder nicht mag. Die Methode der Stimmungsanalyse wird verwendet, um Meinungen, Gefühle und Emotionen aus Texten zu extrahieren. So können auch Phrasen in Produktbeschreibungen, wie sie etwa auf der Website von Amazon zu finden sind, evaluiert werden. Klinger stellte auch seine aktuelle Forschung vor, die sich damit beschäftigt, dieses Modell sprachübergreifend zu optimieren.

Markus Bundschus (Roche Diagnostics GmbH, Penzberg) skizzierte das Thema „*Text and data mining @Roche: an industry perspective*“. In den Lebenswissenschaften gibt es eine Überfülle an Information (allein die Literaturdatenbank PubMed verzeichnet 25 Millionen Zeitschriftenartikeln), sodass es für einzelne Forschende faktisch unmöglich ist, dieses Literaturangebot ohne entsprechende Hilfsinstrumente – automatische Verarbeitung, Filterung, Analyse und Visualisierung – für die eigene Forschung zu nutzen. Bundschus zeigte anhand *Utopia Documents@Roche* ein Konzept, wie Roche das Ziel, wissenschaftliche Mitarbeiterinnen und Mitarbeiter bei ihrer Forschungstätigkeit zu unterstützen, mittels Data Mining erreichen will.

„*Access to knowledge: Text mining and information extraction in the German National Library*“ war das Thema des Vortrags von **Reinhard Altenhöner** (Deutsche Nationalbibliothek, Frankfurt am Main), in dem er den Zugang zu Text und Data Mining an der Deutschen Nationalbibliothek vorstellte. Jährlich gelangen mehr als 400.000 Online-Quellen an die Bibliothek (bei jährlich 800.000 einlangenden gedruckten Medien). Die Katalogisierung und inhaltliche Erschließung dieser Medien erfolgt seit 2010 nicht mehr intellektuell, sondern durch den Einsatz automatisierter Verfahren.

Zum Abschluss des ersten Konferenztages sprach **Martin Hofmann-Apitius** (Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, St. Augustin) über „*Modelling hypothetical knowledge: Capturing and representing scientific speculation in text*“. In wissenschaftlichen Publikationen finden sich immer wieder spekulative Aussagen, die das Ziel verfolgen, weitere Forschungen zum vorliegenden Thema anzustoßen. In jüngster Zeit finden diese spekulativen Aussagen in wissenschaftlichen Texten zunehmend das Interesse von Forschenden, weil eine systematische Analyse dieser Aussagen dazu beitragen kann, sie in nachprüfbare Hypothesen zu verwandeln. Am Beispiel der Alzheimer-Krankheit wurde dargelegt, wie mit einem automatisierten Verfahren (*HypothesisFinder*) eine große Bandbreite von wissenschaftlichen Spekulationen über diese Krankheit herausgefiltert werden kann.

Zu Beginn des zweiten Konferenztages gab **Dietrich Rebholz-Schuhmann** (Universität Zürich) eine Keynote über *„Resolving phenotypes to standard representations: a complex task"*. Big Data gibt es im Bereich der molekularen Daten, der Labordaten und der Patientendaten. Der medizinischen Forschung steht somit eine Fülle an Daten zur Verfügung. Medline verzeichnet 25 Millionen Abstracts; 3 Millionen medizinische Fachartikeln sind als Open Access-Publikationen weltweit frei zugänglich; 20.000 Patientenakten entstehen in einem großen Krankenhaus pro Jahr; durchschnittlich werden einem Patienten oder einer Patientin 20 Medikamente während eines Krankenhausaufenthalts verabreicht; 24.000 Gene wurden entdeckt, die jeden Menschen beschreiben. Eine große Herausforderung für die Zukunft ist es, aus dieser Datenfülle (Big Data) konkretes Wissen (Smart Knowledge) zu generieren. Für zukünftige Innovationen im Gesundheitsbereich ist es erforderlich, die Informationen aus den Patientenakten mit den Krankheitsprofilen und auch mit der Fachliteratur zur betreffenden Krankheit zu vergleichen. Mit diesem Ansatz soll es in Zukunft möglich werden, jeweils relevante Diagnoseverfahren vorzuschlagen, die Gefahr von Nebenwirkungen von Medikamenten für einen konkreten Fall anhand der Patientenprofile abschätzen zu können und das Konzept der translationalen Medizin durch die Integration von Patientendaten zu optimieren.

Im Vortrag von **Stefan Rüger** (The Open University, Milton Keynes) über *„Visual mining – interpretating image data"* ging es nicht um Wörter und semantische Verfahren, sondern um Pixel bzw. Töne. Visual Mining wird durch die rasante Weiterentwicklung im Bereich der Speicherkapazitäten und die damit einhergehende Möglichkeit zur Speicherung von Bildern aus unterschiedlichen Quellen in großer Zahl immer bedeutender. Rüger stellte *New Duplicate Detection* vor, ein Verfahren, bei dem auf der Grundlage einer Smartphone-Anwendung wir *Snaptell*, die dazugehörigen Bilder in einer Produktdatenbank gesucht werden können. Diesem Konzept folgend können auch Töne aus Audiodateien in Bilder umgewandelt und mit anderen Bildern in einer Datenbank abgeglichen werden. Eine weitere Anwendungsmöglichkeit für den Einsatz von maschinellem Lernen und maschinellem Sehen besteht in einer automatisierten Identifikation von ästhetisch ansprechenden bzw. weniger ansprechenden Bildern. Kriterien für Bilder, die als schön empfunden werden, sind etwa Einfachheit, großer Kontrast, Goldener Schnitt und Helligkeit. Durch Nutzung dieser Anwendung könnten Fotokameras in Zukunft so konzipiert werden, dass fotografierende Personen bei der Erstellung perfekter Bilder unterstützt werden.

Lars Juhl Jensen (Universität Kopenhagen) widmete sich dem Thema *„Pragmatic text mining: From literature to electronic health records"*. Text Mining

wird zunehmend ein wichtiges Instrument für das Data Mining in der Medizin. Die Fachliteratur ist eine riesige Datenquelle, deren Inhalte zumeist in den Datenbanken nicht strukturiert erfasst werden; dies gilt etwa für Abbildungen und Grafiken in den Texten. Auch elektronische Gesundheitsakten bilden eine weitere, bisher nur wenig genutzte Datenquelle, die etwa dafür genutzt werden könnte, um unbekannte Zusammenhänge von Krankheiten zu entdecken oder um Medikamente nach deren Zulassung zu verbessern. Juhl-Jensen wies darauf hin, dass in Dänemark alle Patientendaten in einer zentralen Datenbank erfasst werden. Die Vollständigkeit der Datenbank ist dadurch gewährleistet, dass eine Bezahlung für die ärztlichen Leistungen nur dann erfolgt, wenn die entsprechenden Daten abgeliefert wurden.

Anstelle von Juliane Fluck (Fraunhofer-Institut für Algorithmen und Wissenschaftliches Rechnen SCAI, St. Augustin), die an der Konferenzteilnahme kurzfristig verhindert war, übernahm **Martin Hofmann-Apitius** die Präsentation des Vortrags *„BELIEF – a semiautomatic workflow for BEL network creation"*. Um Informationen aus Texten und anderen Datenquellen entnehmen zu können, wurde mit *BEL* (*Biological Expression Language*) ein Verfahren entwickelt, bei dem mittels *Named Entity Recognition* und ausgewählten Wörterbüchern die relevante Terminologie, etwa zu Genen, Proteinen, Krankheiten, Zellen und Gewebe, automatisch ausgewählt und die Beziehung zwischen den Einheiten maschinenlesbar dargestellt wird.

Philipp Daumke (Averbis AG, Freiburg) sprach über *„Large Patent Classification at the European Patent Office"*. Die Firma Averbis und das europäische Patentamt kooperieren bei der Entwicklung eines Verfahrens für Patent Mining. Ziel ist es, die Vorklassifikation von Patentanmeldungen dadurch zu optimieren, dass eingehende Anmeldungen einer oder mehrerer von 1.500 internen EPO-Klassen zugewiesen werden. Zu diesem Zweck wurde ein Algorithmus entwickelt, der für 650.000 Patentanmeldungen der Jahre 2005 bis 2013 testweise angewendet wurde. Besonderes Augenmerk wurde auf die Mehrsprachigkeit gelegt, weil Patentanmeldungen beim Europäischen Patentamt in englischer, französischer und deutscher Sprache erfolgen können.

Stefan Geissler (TEMIS Germany, Heidelberg) und **Matthias Leybold** (Deloitte Consulting AG) widmeten sich dem Thema *„Text Mining and Compliance – Supporting access to complex regulatory legislation by natural language processing"*. Text Mining wird auch zur Strukturierung von Dokumenten über steuerliche Vorgaben eingesetzt werden. Das *FACTA*-Abkommens (*Foreign Account Tax Compliance Act*), das zur Verhinderung von Steuerflucht beitragen soll, führte zu vielen bi-nationalen Abkommen und trug dadurch dazu

bei, das eine enorm große Textmenge entstanden ist. Deshalb haben TEMIS und Deloitte eine Datenbank konzipiert, in der die Sammlung aller einschlägigen Dokumente, deren strukturierte Ablage sowie Aufbereitung für interessierte Steuerexpertinnen und -experten erfolgt.

Der letzte Vortrag wurde von **Anton Heijs** (Datasciencesets, Gouda) über *„Impact of developments in big data analytics for new use cases"* gehalten. Neben Fachliteratur werden in der Medizin auch Tabellen, Bilder und Patientenakten in digitaler Form genutzt. Die Analyse dieser Quellen soll zu neuen Erkenntnissen und damit zu einer Reduktion der Kosten im Gesundheitswesen beitragen. Die Nutzung von Big Data dient vor allem dem Erkennen von Mustern und in der Folge der Entwicklung von Modellen, die dann praktisch angewendet werden können. Derzeit wird dieser erfolgsversprechende Weg, durch eine flächendeckende Analyse von Big Data zu Innovationen zu gelangen, durch das Fehlen dauerhafter Speichermöglichkeiten sowie durch den Einsatz unterschiedlicher Formate beeinträchtigt. Dennoch ist absehbar, dass gerade in der Medizin und den Lebenswissenschaften die Nutzung von Big Data immense Auswirkungen auf die Entwicklung dieser Disziplinen haben wird, auch wenn noch viele komplexe Herausforderungen in naher Zukunft zu bewältigen sein werden.

Zusammenfassend ist festzuhalten, dass die Kölner Konferenz *„From Big Data to Smart Knowledge – Text and Data Mining in Science and Economy"* einen ausgezeichneten Einblick in die Problematik von Big Data, insbesondere aus der Sicht der Natur-, Ingenieur-, Lebens- und Wirtschaftswissenschaften, geboten hat. Besonders interessant erwies sich die Tatsache, dass das Thema überwiegend aus der Perspektive von Forschenden dargestellt worden ist, auch wenn dabei der Aspekt, welche Aufgaben Bibliotheken und Informationseinrichtungen im Kontext von Forschungsdaten in Zukunft übernehmen sollen, etwas ins Hintertreffen geraten ist.

Die Vortragsfolien der meisten Vorträge stehen auf der Konferenz-Website zur Verfügung; von vielen Vorträgen sind auch Videos auf YouTube abrufbar: www.textminingconference.de

Mag. Bruno Bauer
Universitätsbibliothek der Medizinischen Universität Wien
Währinger Gürtel 18–20, 1097 Wien
E-Mail: bruno.bauer@meduniwien.ac.at