„Data scientist is the sexiest job in the 21st century"

Bericht über die 2. European Conference on Data Analysis mit integriertem Workshop on Classification and Subject Indexing in Library and Information Science (LIS 2014) an der Jacobs University Bremen

Frank Scholze und Michael Mönnich

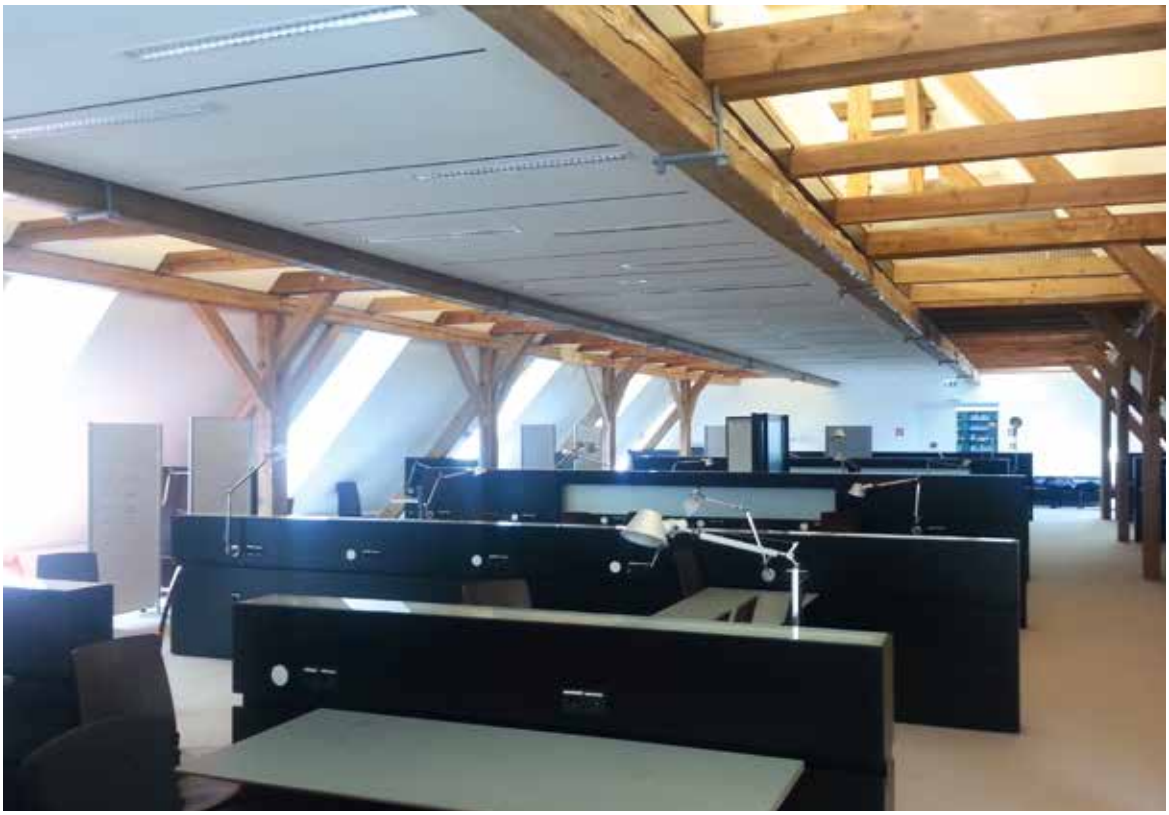
Bibliothek der Jacobs University Bremen

Die Jahrestagung der Deutschen Gesellschaft für Klassifikation (GfKl) fand als 2. European Conference on Data Analysis (ECDA) vom 2. Juli bis 4. Juli 2014 an der Jacobs University Bremen statt[1]. Sie wurde gemeinsam mit der Italian Statistical Society Classification and Data Analysis Group (SIS-Cladag), Vereniging voor Ordinatie en Classificatie (VOC), Sekcja Klasyfikacji i Analizy Danych PTS (SKAD), und der International Association for Statistical Computing (IASC) organisiert und durchgeführt. „Data scientist is the sexiest job in the 21st century" lautete das inoffizielle Motto der Konferenz – ein Zitat aus einem Artikel von Thomas H. Davenport in der Harvard Business Review.[2] Obwohl auch bei der ECDA *Big Data* als Schlagwort im Vordergrund stand, wurde wiederholt festgestellt, dass der Begriff nur dynamisch zu verstehen ist für Datenmengen und Komplexitätsgrade, die sich nicht mit herkömmlichen Verfahren und Methoden analysieren und bearbeiten lassen. Doch diese Grenzen verschieben sich laufend – und wo konnte man dies besser feststellen als bei einer Tagung, auf der neue Forschungsergebnisse vorgestellt wurden, um Daten schneller und effizienter zu analysieren, zu klassifizieren und zu indexieren. So setzte bereits der erste Keynote-Vortrag von Themis Plapanas (Paris Descartes University) den Grundton, in dem er neue Methoden des adaptiven Indexierens von Milliarden von Datenreihen aus dem Bereich der Lebenswissenschaften vorstellte.

Die Konferenz wurde eröffnet mit der Begrüßung durch Katja Wind, der Präsidentin der Jacobs University Bremen. Sie stellte ihre Einrichtung als internationale Universität (75 % der Studierenden sind nicht aus Deutschland) mit den Schwerpunkten Gesundheit, Mobilität und Diversität vor. In einer ehemaligen Kaserne im Westen von Bremen untergebracht, bietet die Jacobs University Studierenden und Dozenten alle Annehmlichkeiten einer angelsächsischen Campus-Universität. Ingo Kramer, Präsident der Bundesvereinigung der Deutschen Arbeitgeberverbände, hob die Bedeutung von Large Data Management hervor, besonders auch unter dem Aspekt des drohenden Mangels an Fachkräf-

1 http://ecda2014.eu/

2 http://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/ar/1



Arbeitsplätze in der Bibliothek

ten. Berthold Lausen, Präsident der Gesellschaft für Klassifikation betonte die Interdisziplinarität des Datenmanagements und kündigte an, die Gesellschaft umzubenennen in Gesellschaft für Datenwissenschaft. Das Tagungsprogramm bestand aus über 150 Vorträgen, rund 200 Teilnehmer waren dem Ruf nach Bremen gefolgt.

Eine Gruppe von Bibliotheks- und Informationswissenschaftlern hielt auch dieses Jahr wieder einen Workshop zu Fragen der Inhaltserschließung ab. Die zeitliche und sprachliche Integration in die Hauptkonferenz waren inzwischen schon gut geübte Routine, so dass in den Pausen und während der parallel stattfindenden Themenblöcke ausreichend Zeit war für den Austausch zwischen Bibliothekaren, Informatikern, Mathematikern, Wirtschafts- und Informationswissenschaftlern.

Der erste Tag des LIS-Workshops begann dann auch mit einem Vortrag von Leon Burkard, einem Wirtschaftsinformatiker vom Lehrstuhl Informationsdienste und elektronische Märkte des Karlsruher Instituts für Technologie (KIT). Er stellte eine Middleware vor (LitObject), die das kooperative Verwalten von Literatur und Publikationen über mehrere Literaturverwaltungsprogramme hinweg ermöglichen soll. Für Zotero ist dies bereits realisiert, Mendeley soll folgen. Burkard warb dafür, die Middleware möglichst breit zu erproben, um ein gutes Feedback für die Weiterentwicklung zu erhalten.

Wie breit gefächert die Themenstellungen derzeit im Bereich Inhaltserschließung sind, machte der zweite Vortrag von Michael Kleineberg vom Institut für Bibliotheks- und Informationswissenschaft der Humboldt Universität zu Berlin deutlich. Er stellte auf der Grundlage von Jürgen Habermas' Formalpragmatik ein konzeptionelles Modell des expliziten und impliziten Wissens vor, bei dem auch das implizite Wissen (Know-how) in einer Form von Kontexterschließung nutzbar gemacht werden kann.

Andreas Kempf vom GESIS Leibniz-Institut für Sozialwissenschaften stellte ein konkretes Vorhaben zur Disambiguierung von Autoren in Sowiport vor. Dieses basiert auf der Verteilung von Erschließungselementen aus dem Thesaurus Sozialwissenschaften (TheSoz) über einen längeren Zeitraum hinweg.

Den ersten Tag des Workshops beschloss ein Besuch an der Staats-und Universitätsbibliothek Bremen. Bei einer Führung durch die Lesesäle und Schulungsräume konnten die Teilnehmer der Tagung einen authentischen Eindruck einer Massenbibliothek mit über 40.000 aktiven Nutzern und rund 3,5 Millionen Medieneinheiten gewinnen.

Frank Seeger, Leiter des Bereiches Bibliotheksdienste der ekz in Reutlingen, gab am zweiten Tag einen Überblick über die in den öffentlichen Bibliotheken im deutschsprachigen Raum verwendeten Klassifikationssysteme. Die älteste ist die 1956 eingeführte Allgemeine Systematik für Öffentliche Bibliotheken (ASB), die 1999 komplett

Pascal Siegers, GESIS in Köln

überarbeitet wurde. Allerdings fanden seit der Schließung des Deutschen Bibliotheksinstituts im Jahre 2004 keine Überarbeitungen mehr statt, bis auf die Teile Technik, Naturwissenschaften und Medizin, die 2014 revisioniert wurden. Weit verbreitet ist auch die 1961 in der DDR entwickelte Klassifikation für Allgemeinbibliotheken (KAB), bei der für das Jahr 2015 eine Onlineversion geplant ist und alle Änderungen mit der ASB synchronisiert werden. Die jüngste Klassifikation ist dieSystematik für Bibliotheken (SfB), 1978 an der Stadtbibliothek Hannover auf der Basis der Systematik der Amerika-Gedenkbibliothek in Berlin entwickelt und bis heute vorbildlich gepflegt. Daneben gibt es noch die 1966 entwickelte Systematik der Stadtbibliothek Duisburg (SSD), die zuletzt 2001 aktualisiert wurde.

Clemens Düpmeier, Leiter des Bereiches Webbasierte Informationssysteme am Institut für Angewandte Informatik des Karlsruher Instituts für Technologie (KIT) referierte über ein technisches Thema: „Storing and Analyzing Bibliographic Metadata with Elastic Search". Die Schnittstellen des World Wide Web zu den Benutzern haben sich in den letzten Jahren stark verändert, weg von komplizierten aus mehreren Feldern bestehenden Eingabemasken hin zu natürlich sprachlichen und einfach zu bedienenden Oberflächen. Diese Anforderung lässt sich am besten mit moderner Suchmaschinentechnologien erfüllen. Ein Beispiel hierfür ist das Portal für die Technikfolgenabschätzung openTA[3], das derzeit im Rahmen eines DFG Projektes in Karlsruhe entwickelt wird und bei dem ElasticSearch zum Einsatz kommt. ElasticSearch basiert wie auch das bekanntere Produkt Solr auf der Lucene-Suchmaschinen-Engine und bietet wie dieses einen dokumentorientierten und strukturierten Suchindex, die einfache Skalierbarkeit und eine Multi-Server-Unterstützung. Darüber hinaus besitzt ElasticSearch eine optimierte Softwarearchitektur, eine starke vor allem aus der Industrie getragene Community und unterstützt die Suche über hierarchische Dokumentstrukturen (Eltern-Kind-Beziehungen). Eine weitere Besonderheit von ElasticSearch sind die so genannten percolators, die eine Analyse von Dokumentinhalten anhand vorgefertigter Abfragen ermöglichen.

Im Vortrag von Tanja Friedrich und Pascal Siegers – beide bei GESIS in Köln beschäftigt – ging es um die inhaltliche Erschließung von Umfragedaten. Alle Umfragen sind über Sowiport erschlossen, jede Umfrage hat einen Satz von Metadaten. Die Nutzer von Sowiport sind weniger an den kompletten Umfragetexten interessiert, sondern steigen eher über die sozialwissenschaftlichen Konstrukte („Bildung") in die Recherche ein. Um diese Anforderungen zu erfüllen, folgen die Referenten dem von Sara Shatford beschriebenen Ansatz[4] der „Ofness and Aboutness" und zerlegen die Gesamtheit der soziologischen Begriffe in vier Hauptklassen und jeweils 5-10 abgeleitete Begriffe, mit denen die Umfragen dann getaggt werden.

Der „Bibliographic report" für das Jahr 2013 wurde dieses Jahr zum ersten Mal gemeinsam von Bernd Lorenz (Fachhochschule für öffentliche Verwaltung und Rechtspflege in Bayern, München) und Michael Franke (Universitätsbibliothek der FU Berlin) vorgetragen und umfasste neben dem bewährten Überblick über wichtige Publikationen aus dem Bereich der Klassifikation nun auch eine Auswahl an interessanten und wichtigen Webquellen zum Thema (z.B. den Dewey-Browser von BASE oder das Basel Register of Thesauri, Ontologies and Classifications).[5]

Neben dem Vortragsprogramm6 blieb viel Raum für Gespräche und Diskussionen unter den Teilnehmern. Dieselbe gelungene Kombination soll auch der nächste LIS-Workshop bieten. Dieser findet – dann im Rahmen der 3. European Conference on Data Analysis – am 2./3. September 2015 an der University of Essex in Colchester statt. ▮

3 http://www.openta.net

4 Sara Shatford, „Analyzing the Subject of a Picture: A Theoretical Approach", In: Cataloging & Classification Quarterly Volume 6, Issue 3, 1986, pages 39-62, DOI: 10.1300/J104v06n03_04

5 http://tinyurl.com/nl7jpe5

6 Die Folien sämtlicher Präsentationen können über die KIT-Bibliothek (http://tinyurl.com/nnov8pc) abgerufen werden:

Prof. Dr. Michael Mönnich
KIT-Bibliothek
Leiter der Abteilung Benutzung
Straße am Forum 2
76131 Karlsruhe
michael.moennich@kit.edu

Frank Scholze
Karlsruher Institut für Technologie
Direktor der KIT-Bibliothek