

## Tagungsberichte

### Klassifikation international

#### Bericht über die European Conference on Data Analysis (ECDA) mit integriertem Workshop on Classification and Subject Indexing in Library and Information Science (LIS'2015) in Colchester

##### Die ECDA 2015

Zum nunmehr dritten Mal richtete die Gesellschaft für Klassifikation (GfKl) gemeinsam mit internationalen Partnerverbänden – der britischen Gesellschaft für Klassifikation und der Sektion für Klassifikation innerhalb der polnischen Gesellschaft für Statistik – eine europäische Konferenz aus, die „European Conference on Data Analysis“ (ECDA). Über 130 Wissenschaftlerinnen und Wissenschaftler aus 19 Ländern kamen vom 2. bis 4. September 2015 an der University of Essex in Colchester (Großbritannien) zusammen. „Data Science: Foundations, Methods and Applications“ lautete das Thema der Konferenz. Es wurde deutlich, dass der Hype um *Big Data* vorbei ist und der seit 1996 im Konferenztitel<sup>1</sup> gebrauchte Begriff *Data Science* treffender ist, um die Extraktion von Wissen aus Daten zu benennen.



Abb. 1: Der Tagungsort – das neue Gebäude der Essex Business School (Foto: Wiesenmüller)

1 C. Hayashi u.a. (Hg.): Data science, classification, and related methods. Proceedings of the Fifth Conference of the International Federation of Classification Societies (IFCS-96), Kobe, Japan, March 27–30, 1996, Tokyo u.a.: Springer, 1998.

Der Haupt-Campus der University of Essex liegt im Grünen, etwas außerhalb des Stadtzentrums von Colchester. Die Universität wurde 1963 gegründet und hat derzeit ca. 12.000 Studierende – mehr als 5.000 davon kommen nicht aus Großbritannien. Quadratische Höfe und turmartige Hochhäuser prägen das architektonische Bild der Universität; vorherrschend ist Beton. Die Tagung fand jedoch in der erst vor Kurzem eröffneten, nachhaltig in Holzbauweise errichteten *Essex Business School* statt (Abb. 1 und 2). Diese bot einen beeindruckenden und angenehmen Rahmen – wenn auch die Technik des vollautomatisierten Gebäudes noch nicht in allen Details funktionierte.

Auch in diesem Jahr hielt eine Gruppe von Bibliotheks- und Informationswissenschaftler/innen im Rahmen der Konferenz einen Workshop zu Fragen der Klassifikation und Inhaltserschließung ab, den „Workshop on Classification and Subject Indexing in Library and Information Science“ (LIS'2015). Die zeitliche und sprachliche Integration des LIS-Workshops in die Hauptkonferenz sind inzwischen schon gut geübte Routine. Im Zeitraster der Tagung gibt es bestimmte Phasen, in denen nur ein bzw. zwei Tracks parallel laufen (die sogenannten *Plenaries* bzw. *Semi-Plenaries*). Diese sind für alle Teilnehmerinnen und Teilnehmer gedacht und ermöglichen den Kontakt zwischen den unterschiedlichen Disziplinen, die sich in der restlichen Zeit auf verschiedene Sektionen verteilen. Auch die Pausen, in denen wir im Foyer verköstigt wurden, boten ausreichend Zeit für den Austausch zwischen Kolleginnen und Kollegen mit unterschiedlichem Background. Vertreten sind auf der ECDA insbesondere die Informatik, die Mathematik und die Wirtschaftswissenschaften.

## Teil 1 des LIS-Workshops

Der erste Tag des LIS-Workshops begann mit einem Vortrag von Andreas Kempf von der Deutschen Zentralbibliothek für Wirtschaftswissenschaften (ZBW), der seine aktuelle Forschungsarbeit vorstellte:<sup>2</sup> ein iteratives, semiautomatisches Mapping des Standard-Thesaurus Wirtschaft (STW) und der Klassifikation des „Journal of Economic Literature“ (JEL). Der Ansatz basiert auf einer Abbildung der Fächerklassen des STW. Bis Mitte 2016 soll eine produktive Umgebung für den Teilbereich Volkswirtschaftslehre vorliegen.

Im Anschluss daran stellte Martin Mehlberg von der Technischen Informationsbibliothek Hannover (TIB) die Pläne zur Weiterentwicklung des „Thesaurus Engineering und Management“ (TEMA) in Kooperation mit der WTI Frankfurt eG vor.<sup>3</sup> Als Vorarbeiten wurde TEMA auf die Averbis Terminologieplattform migriert und mit Begriffen aus der Gemeinsamen Normdatei (GND) im Bereich Ingenieurwissenschaft angereichert. Angestrebt werden nun eine kontinuierliche Aktualisierung des Thesaurus mittels Textmining-Methoden sowie eine semiautomatische englische Übersetzung. Dies wird anhand des thematischen Testbereiches Elektromobilität bis Ende 2016 untersucht.

---

2 Building the bridge – mapping different knowledge organization systems in economics. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-496582> (11.10.2015).

3 Towards a comprehensive knowledge organisation system for the engineering domain. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-499293> (26.10.2015).



Abb. 2: Der Wintergarten in der Essex Business School (Foto: Wiesenmüller)

Nach der Mittagspause zeigte Heidrun Wiesenmüller von der Hochschule der Medien in Stuttgart die Implikationen des neuen Regelwerks „Resource Description and Access“ (RDA) für die inhaltliche Erschließung auf.<sup>4</sup> Im ersten Teil wurde erläutert, wie sich die Sacherschließung in die unterschiedlichen FR-Modelle<sup>5</sup> und in RDA einfügt. Seit einer Regelwerksänderung vom April 2015 ist klar, dass RDA nicht den Anspruch hat, die speziellen Regelwerke und Instrumentarien für die inhaltliche Erschließung zu ersetzen. Im zweiten Teil wurden Möglichkeiten und Grenzen der Nutzung von RDA für die Sacherschließung im deutschsprachigen Raum betrachtet. Deutlich wurde dabei auch, dass Formal- und Sacherschließung aufgrund der RDA-Einführung weitaus enger zusammenarbeiten werden als bisher.

Zum Abschluss des ersten Tages präsentierte Friedrich Summann von der Universitätsbibliothek Bielefeld, wie Sacherschließungsdaten beim Indexieren von Open-Access-Repositorien im Rahmen der Bielefeld Academic Search Engine (BASE) erfasst und ausgewertet werden.<sup>6</sup> Eine zentrale Problematik ist dabei die Heterogenität der global verteilt vorliegenden Informationen. Konkret werden mit computerlinguistischen Methoden durch Konkordanzen zu vorliegenden Fachklassifikationen

4 Subject cataloguing in an RDA framework – strategies and practical experience from Germany. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-494159> (11.10.2015).

5 „FR“ steht für „Functional Requirements“; das wichtigste FR-Modell ist FRBR (Functional Requirements for Bibliographic Records).

6 The role of classification information in Open Access repositories – current status and future directions. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-494574> (11.10.2015).

einheitlich DDC-Klassifikationen der ersten drei Hierarchieebenen für Dokumente vergeben, die in englischer oder deutscher Sprache vorliegen. Von rund 78 Millionen in BASE indixierten Dokumenten sind rund 10 Millionen auf diese Weise erschlossen. Eine Weiterentwicklung dieses vielversprechenden Erschließungsansatzes erscheint wünschenswert.

Nach dem Vortragsprogramm gab es für die Teilnehmenden des LIS-Workshops eine Führung durch die Hauptbibliothek der Universität (Albert Sloman Library), die gerade einen Erweiterungsbau erhalten hat (Abb. 3). Für deutschsprachige Bibliothekarinnen und Bibliothekare ungewohnt sind die Benutzungsregeln: So beträgt die normale Ausleihfrist für *undergraduates* nur zwei Wochen, und stark nachgefragte Literatur ist oft nur per Kurzausleihe für wenige Stunden verfügbar. Den Tagesausklang bildete ein gemeinsames Abendessen.



Abb. 3: Impression aus der Albert Sloman Library  
(Foto: Wiesenmüller)

## Teil 2 des LIS-Workshops

Der Vormittag des zweiten Tages begann mit einem Vortrag von René Hackl-Sommer vom FIZ Karlsruhe.<sup>7</sup> In seinem Vortrag erläuterte er den aktuellen Stand eines Werkzeugs zur Analyse eines wichtigen Textteils von Patentanmeldungen, der Patentansprüche (engl. *patent claims*). Dieser Textteil ist oft umfangreich und beschreibt zum einen die kennzeichnenden Merkmale, die die Erfindung patentwürdig machen, und zum anderen, welcher Schutz im Einzelnen beansprucht wird. Für die Recherche in Patentschriften ist eine Unterscheidung zwischen unabhängigen und abhängigen Ansprüchen wichtig, ebenso wie deren Beziehungen untereinander. Das am FIZ entwickelte Werkzeug nutzt die Tatsache, dass Patente als juristische Texte stark formalisiert sind. Mit Hilfe eines regelbasierten Textmining-Algorithmus kann es die gewünschten Merkmale mit guter Genauigkeit extrahieren. Die Ergebnisse sollen schon bald für eine verbesserte Patentrecherche am FIZ praktisch eingesetzt werden.

Der zweite Vortrag wurde von Wolfram Sperber, der ebenfalls am FIZ Karlsruhe beschäftigt ist, gehalten.<sup>8</sup> Er berichtete vom Projekt „MathSearch“, das es sich zur Aufgabe gemacht hat, die Suche in mathematischen Volltexten bis auf die Ebene einzelner Formeln zu

7 Patent claim structure recognition. Die Folien werden in Kürze bei der KIT-Bibliothek bereitgestellt; abrufbar unter <http://services.bibliothek.kit.edu/primo/start.php?query=lis+2015&sort=date&tab=evastar> (11.10.2015).

8 SMGloM – a semantic mathematical glossary of the next generation. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-495014> (11.10.2015).

ermöglichen. Die Schwierigkeit ist dabei, dass die in den Formeln verwendeten Symbole zwar als Zeichen enthalten sind, ihre eigentliche Bedeutung aber stark vom Kontext abhängig ist. Im Projekt sollen mathematische Formeln in Texten analysiert und durch eine Auswertung des umgebenden Texts automatisch disambiguiert werden. Dazu wurde prototypisch ein Glossar entwickelt, das die relevanten mathematischen Konzepte in verschiedenen Sprachen beschreibt und die verwendeten Symbole dokumentiert. Die Entwicklung eines solchen Glossars ist sehr aufwendig und soll künftig von der mathematischen Community als Online-Collaboration fortgeführt werden.

In der abschließenden Nachmittags-Session berichtete zunächst Uma Balakrishnan von der Verbundzentrale des Gemeinsamen Bibliotheksverbunds (GBV) über den derzeitigen Stand des Mapping-Werkzeugs Cocoda.<sup>9</sup> Dabei wurde auch ein Prototyp demonstriert. Cocoda ist ein Open-Source-Werkzeug zum teilautomatischen Erstellen von Konkordanzen. Das Web-Tool liest Daten aus unterschiedlichen Quellen ein, stellt die Informationen aus Quell- und Zielsystem übersichtlich dar und macht Vorschläge für das Mapping.

Danach beschäftigten wir uns mit einem Beispiel aus den Niederlanden: Christian Wartena von der Hochschule Hannover stellte seine Untersuchungen an Dokumenten aus dem Niederländischen Parlament vor.<sup>10</sup> Da die Erschließungsarbeit durch den Informationsdienst des Parlaments intellektuell nicht mehr zu leisten ist, ist man an automatischer Erschließung interessiert. Der Einsatz eines kommerziellen Systems zur automatischen Klassifizierung lieferte jedoch enttäuschende Ergebnisse. Das Ziel ist es nun, über computerlinguistische Methoden den Parlaments-Thesaurus mit zusätzlichen Synonymen anzureichern, um eine bessere Basis für maschinelle Methoden zu erhalten. Das Grundprinzip sind paradigmatische Vergleiche: Kommen zwei Wörter häufig im selben Textzusammenhang vor, so ist anzunehmen, dass sie semantisch sehr ähnlich sind. Das Auffinden von Synonym-Paaren funktioniert auf diesem Weg recht gut. Hingegen erweist es sich als schwierig, für beliebige Paare von Wörtern zu entscheiden, ob diese zum selben Konzept gehören oder nicht.

Den traditionellen Abschluss des Programms bildete der bibliographische Bericht: Gerald Peichl von der Universität St. Gallen stellte eine Auswahl aktueller Literatur zu Themen der Sacherschließung vor.<sup>11</sup> Ergänzt wurde dies durch den „Online-Report“ von Michael Franke-Meier (Freie Universität Berlin), der in Form einer Prezi-Präsentation interessante Tools und Features aus dem Web demonstrierte.<sup>12</sup> Dazu gehören u.a. ein Unterstützungstool für die Fachreferatsarbeit, das im Rahmen von „malibu“ (Mannheim library utilities) an der UB Mannheim entstanden ist:<sup>13</sup> Bei Eingabe einer ISBN oder einer Identnummer aus einem Bibliotheksverbund ermittelt das System automatisch

---

9 Colibri Concordance Database (Cocoda) – a mapping tool for library classification schemes. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-497791> (11.10.2015).

10 Automatic identification of synonym relations in the Dutch Parliament Thesaurus. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-494163> (11.10.2015).

11 Bibliographic Report 2015. A choice of relevant classification literature. Folien unter <http://nbn-resolving.org/urn:nbn:de:swb:90-494494> (11.10.2015).

12 A choice of nice web-features for subject cataloguing. Prezi-Präsentation unter [https://prezi.com/wtbqfzza1unc/online-report/?utm\\_campaign=share&utm\\_medium=copy](https://prezi.com/wtbqfzza1unc/online-report/?utm_campaign=share&utm_medium=copy) (11.10.2015).

13 Unterstützungstool für die Fachreferatsarbeit: <http://data.bib.uni-mannheim.de/malibu/isbn/suche.html> (11.10.2015). Informationen zu malibu unter <https://github.com/UB-Mannheim/malibu> (11.10.2015).

die bereits vergebenen Sacherschließungsdaten (DDC, RVK, Schlagwörter) in verschiedenen Verbänden sowie weitere Informationen zum Medium. Kurz gezeigt wurde beispielsweise auch das in den Niederlanden an der Koninklijke Bibliotheek entwickelte „Named entity recognition tool for Europeana newspapers“.<sup>14</sup>

Am Abend bestand die Möglichkeit zur Teilnahme an einer Stadtführung mit dem Schwerpunkt auf der römischen Geschichte der Stadt sowie dem Conference Dinner in Colchester Castle (Abb. 4). Aber auch sonst blieb viel Raum für Gespräche und Diskussionen unter den Teilnehmenden.



Abb. 4: Blick auf Colchester Castle (Foto: Wiesenmüller)

## **Ausblick**

Im kommenden Jahr findet die 40. Jahrestagung der GfKI als Teil der Tagung DAGStat in Göttingen statt – allerdings genau parallel zum Leipziger Bibliothekskongress. Damit sich die beiden Veranstaltungen nicht gegenseitig Konkurrenz machen, wurde vereinbart, dass der LIS-Workshop 2016 als Teil des Bibliothekskongresses in Leipzig ausgerichtet wird, und zwar am Mittwoch, 18. März 2016.

14 <https://github.com/kbnlresearch/europeanap-ner> (11.10.2015).

Im Jahr 2017 wird wieder ein Ziel im europäischen Ausland angesteuert: Die „European Conference on Data Analysis“ findet dann – inklusive des LIS-Workshops – in der südwestpolnischen Metropole Wroclaw statt.

*Heidrun Wiesenmüller, Hochschule der Medien*

*Frank Scholze, KIT-Bibliothek*

*Magnus Pfeffer, Hochschule der Medien*

**Zitierfähiger Link (DOI):** <http://dx.doi.org/10.5282/o-bib/2015H4S314-320>