Florian Betz

Automatic indexing at the German National Library (DNB). Experiences and results

Table of contents

1. Chronological summary of efforts on automatic indexing at the DNB

Mandate of the German National Library (DNB)

"to collect in the original, to inventory, to catalogue and bibliographically index, to permanently safeguard and to prepare for use by the general public

a) media works published in Germany from 1913 on and

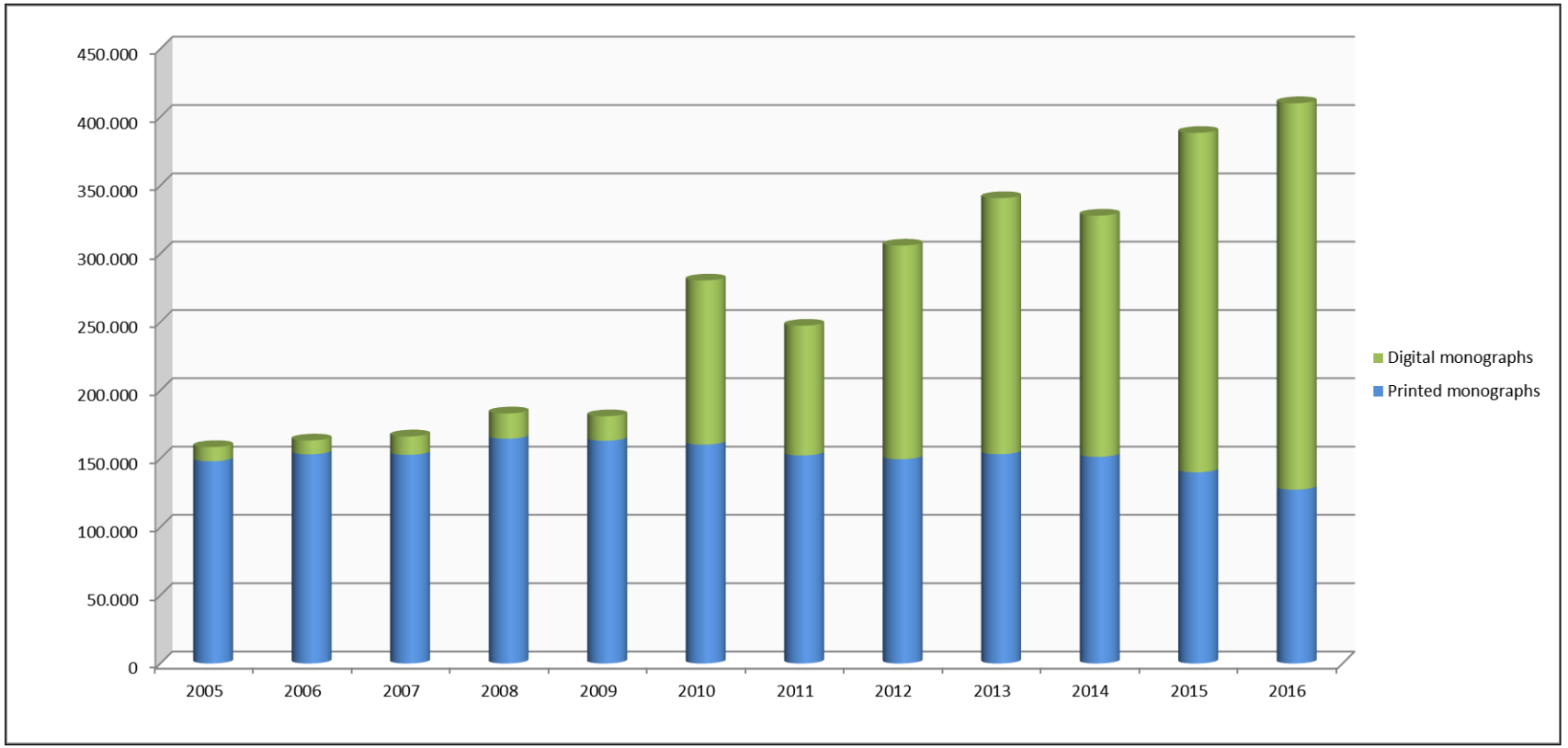
b) German-language media works, translations of German-language media works into other languages and foreign-language media works about Germany published abroad from 1913 on

and to provide central library and national bibliographic service"

Collection of online publications

- 1996-2006: Collecting of online publications but no legal obligation
- 2006: Official collection mandate for online publications by the „Law regarding the German National Library“ (DNBG)

Background I:
Increase of incoming publications



Background II: Intellectual subject cataloguing at the DNB

- Subject categorisation:
  - assignment of about 100 DNB subject categories based on DDC numbers (level of Hundreds, e.g. 150 Psychology, 330 Economics)
  - obligatory for all publications
  - used as classificatory arrangement for the Deutsche Nationalbibliografie
- Classificatory indexing: assignment of full DDC numbers
- Subject indexing:
  - assignment of subject headings of the Integrated Authority File (GND)
  - syntactical indexing according to the RSWK (Rules for Subject Cataloguing)

Example: Title announcement in the Deutsche Nationalbibliografie (series A)

<740> XA-DE-BW
http://d-nb.info/112033005X ∞ 16,N49
Hellweg, Marion: Sweet Living im Scandi Style / Fotos von Shanti Broeng ; Marion Hellweg. - 1. Auflage. - Stuttgart : Lifestyle BusseSeewald, 2017. - 159 Seiten : Illustrationen ; 29 cm . - Inhaltstext . - Inhaltsverzeichnis - ISBN 978-3-7724-7444-6 Festeinband : EUR 24.95 (DE), EUR 25.70 (AT), CHF 32.50 (freier Preis) - ISBN 3-7724-7444-6 - EAN 9783772474446
SW: Skandinavien ; Wohnen ; Innenarchitektur
DDC: 747.0948

Automatic cataloguing: How it began?

2010: Introduction of new series „O“ of the Deutsche Nationalbibliografie

- Series O comprises all online publications (formerly media-independently sorted).
- At the same time termination of intellectual subject cataloguing of online publications

The Project PETRUS (2009-2011)

- PETRUS = Process-supporting software for the digital German National Library
- Large inter-divisional project for developing and establishing automatic cataloguing procedures
- Current automatic subject indexing procedures at the DNB originally go back to scenario 4 of this project.
- Subsequently institutionalisation and further development of the different scenarios to productive automatic cataloguing procedures

Further chronology of productive automatic subject cataloguing procedures

- 2012: Start of automatic assignment of DNB subject categories for German- and English-language titles
- 2014: Start of automatic subject indexing of German-language university publications
- 2015: Start of automatic assignment of DDC short numbers for medicine for German- and English-language university publications
- 2017: Start of automatic subject indexing of German-language BoD-titles (without fiction) and of scientific articles
- Latest: Start of automatic subject indexing of print publications by digitised ToCs (university publications; publications outside the publishers‘ booktrade)

2014: Organisational reform

- The Subject area Automatic indexing was prior to 2014 part of the Departement of Subject Cataloguing.
- 2014: Formation of the „Section AEN: Automatic indexing; Online publications“ by centralisation of formerly dispersed subject areas.
- Section AEN comprises automatic subject cataloguing as well as big parts of aquisition and descriptive cataloguing of online publications (last only for periodicals).

DEUTSCHE NATIONAL BIBLIOTHEK

Table of contents

Objectives and principles I

- Method: Computational linguistic approach combined with the use of a dictionary
- Basis: bibliographic metadata; fulltext; digitised ToCs
- Document formats: PDF; Epub
- Terminology: Integrated Authority File (GND)
- Output: currently maximum 10 subject headings per document or all subject headings above a defined threshold
- Software: Averbis GmbH (Freiburg im Breisgau)
  - Averbis Extraction Platform (AEP)
  - Averbis Terminology Platform (ATP)

Objectives and principles II

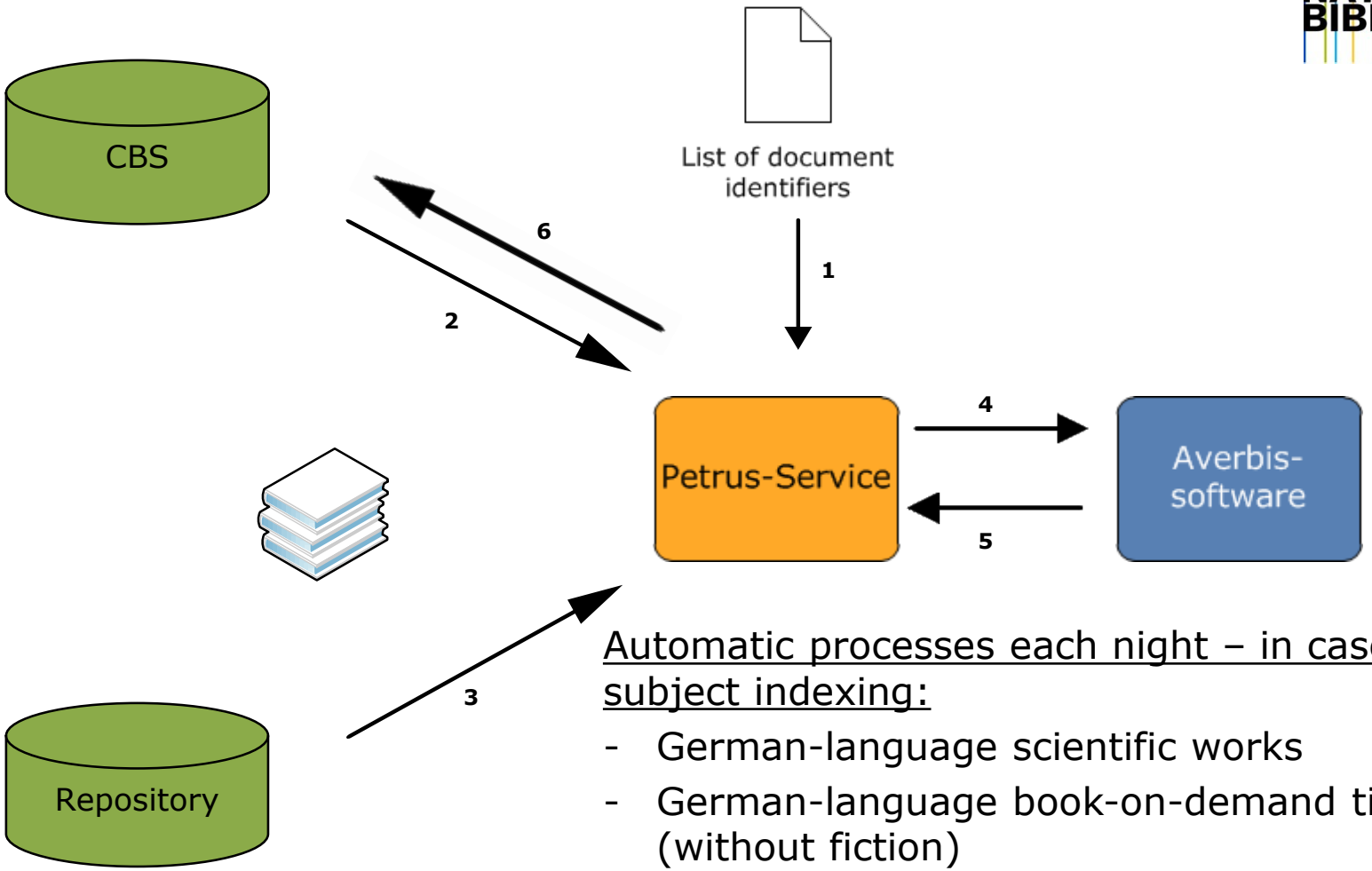
- Goal is a homogeneous coverage with subject headings of the Integrated Authority File (GND) for printed and online publications, by intellectual and automated indexing.
- Dictionary care is indispensable and time-consuming.
- Not in all cases the GND includes the needed concept.
- Automated subject indexing results are of lower quality than the results of intellectual indexing.
- Criterion for the quality of machine-generated subject headings is the usefulness for retrieval purposes.
- Machine-generated subject heading strings are non syntactical.

Integrated Authority File (GND)

- The Integrated Authority File contains headings for descriptive as well as for subject cataloguing (so-called partial stock "s").
- The copy of this partial stock „s" as currently used for automatic indexing counts:

Entity	**Quantity**	
Personal names	373.852	records
Topical headings	185.594	records
Geographical names	207.514	records
Corporate body names	144.224	records
Conferences and Events	12.191	records
Title headings	88.015	records
Total	**1.011.390**	records

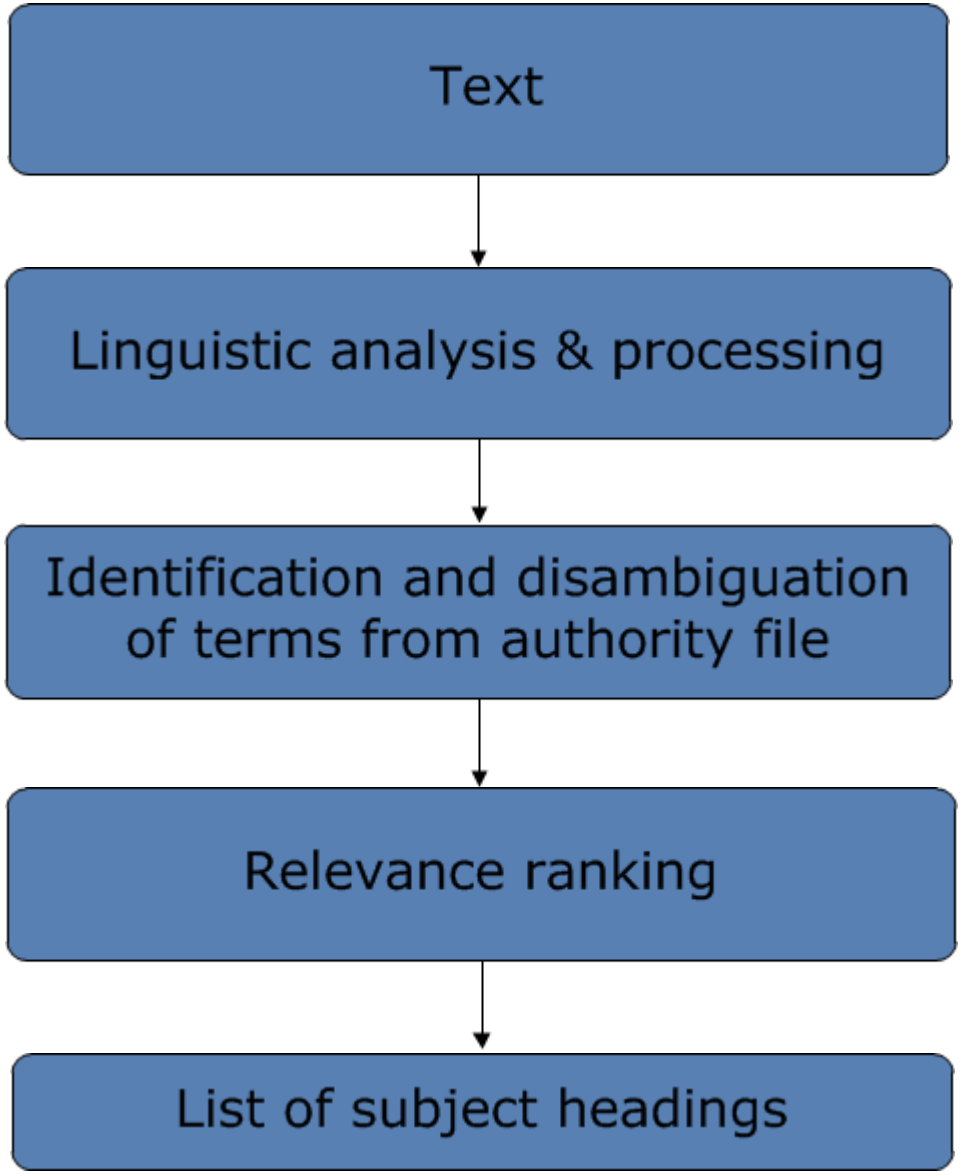
Workflow and Routine



<u>Automatic processes each night – in case of subject indexing:</u>

- German-language scientific works
- German-language book-on-demand titles (without fiction)
- German-language scientific articles
- German-language digitised ToCs of series B and H (print publications)

Averbis Software: processing steps



DEUTSCHE NATIONAL BIBLIOTHEK

Ambiguity

Bank [bank -> Bank]


The Bank of England, established in 1694.

[bank -> Bank <Möbel>]

Averbis Software: Dictionary

- Terminologies are cared on a profile basis.
- Profile = set of single or systematic manipulations of a terminology for adjusting it to the needs of automatic processes
- For systematic and extensive modifications complex filters can be defined using regular expressions.
- Levels of terminology:
  - Hierarchies of concepts (leaf nodes)
  - Concepts
  - Terms
  - Mappingforms

Averbis Software: Configuration

- Configuration = combination of parameter settings for regulating the indexing process
- The used dictionary profile is one parameter of a configuration.

Entry of machine-generated subject cataloguing data in title record

title: Theater ist eine Volkssauna. Politisches Gegenwartstheater aus Finnland in der Tradition von Bertolt Brecht

subject cataloguing entries:

```
5050 792;890;830$Ep$D2015-05-06
5050 700$Ea$Hwbf$D2014-09-24
5051 $LS_WA3_WB38_20160510_de
5100 !118514768!Brecht, Bertolt [Tp1]
5101 !04049716X!Rezeption [Ts1]
5102 !040172430!Finnland [Tg1]
5103 !04046587X!Politisches Theater [Ts1]
5109 (DE-101){DE-101}
5400 [DDC22ger]792.094897
5401 792.09
5403 -T2--4897
5540 [GND]!040597024!Theater [Ts1]$K0,308$D2016-05-24
5540 [GND]!118514768!Brecht, Bertolt [Tp1]$K0,016$D2016-05-24
5540 [GND]!040172430!Finnland [Tg1]$K0,013$D2016-05-24
5540 [GND]!040128997!Drama [Ts1]$K0,007$D2016-05-24
5540 [GND]!041306236!Volkstheater [Ts1]$K0,004$D2016-05-24
5540 [GND]!041402413!Dramatiker [Ts1]$K0,003$D2016-05-24
```

5050 DNB subject categories $E p=from parallel edition| a=deliverer $H origin (wbf=webform) $D date

5051 $L configuration used for automatic subject indexing

510X intellectually assigned GND subject headings

540X intellectually assigned DDC numbers

5540 machine-generated subject headings
$K confidential value $D date

Entry of machine-generated subject cataloguing data in title record

title: Symposium: Computational Chemistry, Physics and Biology

foreign data LCSH:

5560 [LCSH]Universities; Germany; Ulm; Congresses
5560 [LCSH]Computational biology; Congresses
5560 [LCSH]Computational chemistry; Congresses
5560 [LCSH]Food; Congresses
5560 [LCSH]Spectroscopy; Congresses

automatic subject cataloguing entries:

5050 540EmHdnb$K0,958$D2015-10-01
5050 000EaHxmp$D2015-09-30
5051 $KK_A4_03_20140925_de$LS_WA3_WB38_20160510_de
5052 $f500$F0,738$g610$G0,692$D2015-10-01
5540 [GND]!042900913!*Computational chemistry [Ts1]*$K0,295$D2016-05-19
5540 [GND]!040615294!*Ulm [Tg1]*$K0,049$D2016-05-19
5540 [GND]!04045956X!*Physik [Ts1]*$K0,012$D2016-05-19
5540 [GND]!040437930!*Organische Chemie [Ts1]*$K0,006$D2016-05-19

Display of machine-generated subject headings in the DNB catalogue

Link zu diesem Datensatz	http://d-nb.info/1077042000
Titel	Symposium: Computational Chemistry, Physics and Biology / Hans-Ullrich Siehl
Person(en)	Siehl, Hans-Ullrich
Verlag	Ulm : Universität Ulm. Kommunikations- und Informationszentrum
Zeitliche Einordnung	Erscheinungsdatum: 2013
Umfang/Format	Online-Ressource
Persistent Identifier	URN: urn:nbn:de:bsz:289-vts-87752
URL	http://vts.uni-ulm.de/docs/2013/8775/vts_8775.zip (Verlag) (kostenfrei zugänglich)
Sprache(n)	Deutsch (ger)
Anmerkungen	Langzeitarchivierung gewährleistet
Schlagwörter	Lebensmittel ; Quantenchemie ; Physics; Computer programs; Congresses ; Quantum chemistry; Congresses ; Spektroskopie ; Universität Ulm Computational chemistry* ; Ulm* ; Physik* ; Organische Chemie* (*maschinell ermittelt)
Sachgruppe(n)	540 Chemie

subject headings

(*machine-generated)

DEUTSCHE NATIONAL BIBLIOTHEK

Table of contents

1. **Chronological summary of efforts on automatic indexing at the German National Library (DNB)**
2. **Automatic indexing with the German-language Integrated Authority File (GND)**
3. **Development of automatic indexing with the English-language LCSH**

Project MAEN

- MAEN = automatic indexing of English-language online publications
- Initial position:
  - 2010 stop of intellectual indexing of online publications (series O)
  - English-language online publications are not covered by current productive automatic indexing routines with the Integrated Authority File (GND).
  - English-language online publications are by far the second largest group of online publications in the entire stock.
- Project term: 04/2016-06/2018

Objectives and principles

- Subsequent use of the Averbis Software after adjustments and additions for English
- Primary goal: automatic assignment of English-language subject headings taken from a controlled vocabulary
- Use of the Library of Congress Subject Headings (LCSH) as indexing terminology and output terminology

Library of Congress Subject Headings (LCSH)

- Universal English-language subject headings authority file of the Library of Congress
- Approximately 415.000 concepts and 350.000 synonyms
- Structural principle: precoordination
- Low coverage of LCSH by LCC numbers (about 25%)
- For subject indexing used in combination with the Name Authority File (NAF) of the Library of Congress

LCSH as imported terminology in the Averbis Software

- Import: SKOS/RDF-XML dump of the Library of Congress
- Besides preferred and alternative labels creation of additional mappingforms for each of them according to specific rules, for example:

label: Phototropism (Chemistry)
additional mapping forms:
Chemistry Phototropism
Chemistry

Positive indexing example of current development status:

title: Positive Psychology and Change : How Leadership, Collaboration, and Appreciative Inquiry Create Transformational Results

machine-generated LCSH:
Leadership
Change (Psychology)
Management
Appreciative inquiry
Positive psychology
Organizational change

preferred label: Organizational change

text matches with synonym: **Organizational Development**

Negative indexing example of current development status:

title: The contribution of non-formal schools in enhancing the provision of basic education in Kenya

intellectually determined keywords:
Alternative Schools
Disadvantaged Children
Basic Education
Kenya

machine-generated LCSH:
Basic education
Schools
Boys
Girls
Kenya (Southeast Asian People)
Learning
Teachers

First findings

Dictionary care:
Filtering terms with unwanted qualifier-patterns causing systematical errors is possible. – Further example: „Einstein (Horse)“.

Terminology:
Without further entities, especially personal and geographical names, there is an unwanted shift towards false homonymous names as substitutes.

Table of contents

Outlook

- Further development of indexing procedure
- Developing of a procedure and workflow of text structure analysis, recognition and segmentation
- Retrospective indexing of online publications
- Data delivery of machine generated subject heading strings
- Further development and extension of automatic indexing of digitised ToCs taken from print publications
- Project TeMa (= Terminology management for support of intellectual and automatic subject cataloguing)

Thank you for your attention.

Questions?

Further information: DNB/Automatic cataloguing

Literature:

Ulrike Junger (2014) Can Indexing Be Automated? The Example of the Deutsche Nationalbibliothek, Cataloging & Classification Quarterly, 52:1, 102-109, DOI: 10.1080/01639374.2013.854127 / http://dx.doi.org/10.1080/01639374.2013.854127

Contact:

Dr. Florian Betz
Deutsche Nationalbibliothek / German National Library
Automatic indexing; Online Publications
Phone: +49–341–2271–588
mailto:f.betz@dnb.de