Subject Cataloging Policies in the Deutsche Nationalbibliothek

Preparing for the Future

Ulrike Junger
Deutsche Nationalbibliothek

Overview

1. Deutsche Nationalbibliothek – some facts and figures
2. Subject cataloging in DNB
3. Challenges
4. Advancing the concept of subject cataloguing
5. Automatic subject cataloging
6. Reuse of subject data
7. Concluding remarks

Deutsche Nationalbibliothek (DNB) – some facts and figures (I)

- Legal deposit: all German and German-language publications, publications about Germany etc from 1913 (since 2006 including non-physical publications)
- Mandate: Collection, cataloguing, (long-term) archiving and make the media units available to the general public
- Main products:
  - National Bibliography (Catalogue)
  - Gemeinsame Normdatei (GND)

2 sites: Leipzig + Frankfurt

DNB – some facts and figures (II)

- Collection size: 28 million media units
- 860.000 units were added in 2012:
  - 610.000 physical units
    (books, theses, journals, sound carriers, sheet music, standards etc)
  - 250.000 non-physical units
    (online publications, e-journals, e-paper editions, websites etc)
- Catalogue: 12 million records (title data)
- Gemeinsame Normdatei: 10 million records (subject headings, geographical terms, name authorities, corporate bodies etc)

Subject cataloging in DNB

Indexing with GND subject headings

most physical publications from the book publishing trade

Classification with DeweyDecimal Classififcation

for most physical publications

DDC-based subject groups

for all physical publications

Different series of the Deutsche Nationalbibliografie as a structuring principle

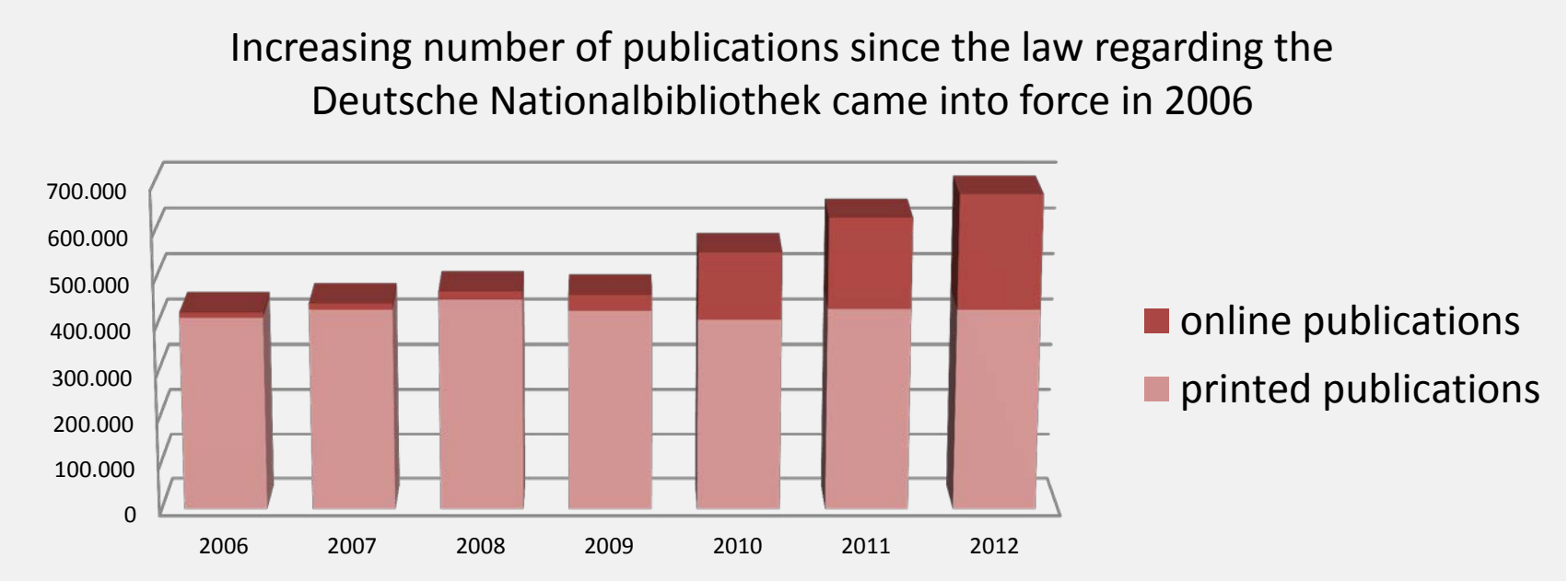
Subject cataloging in DNB

- All online publications are listed in the Series O of the Deutsche Nationalbibliografie
- Intellectual cataloging was stopped in 2010

Challenges

- Changes in publication landscape and structure of DNB‘s collection (increasing number of e-only publications, Web harvesting …)
- Internationalisation of library standards, formats and structures (RDA, BibFrame), boundaries between descriptive and subject cataloging loosen up
- New methods and tools in the web (Linked Data, …), including „webization“ of standards (BIbFrame)
- Demographic changes: between 2011 and 2016 retirement of 20% of the staff in department of subject cataloging

Challenges II



Consequence: Advancing the concept of subject cataloguing is necessary

- Changes in intellectual subject cataloging
- More resources for the development of new procedures and tools
- Different paths can lead to the goal (=good subject access for users)

Advancing the concept for subject cataloging

Major goal:

To index as many publications as possible in an appropriate manner

Further goals:

- Stabilizing the resource situation
- Production of well reusable metadata
- New procedures to maintain the GND authoritiy file

Advancing the concept of subject cataloging
Future directions

Stronger diversification of subject cataloging

- Methods
- Range and depth

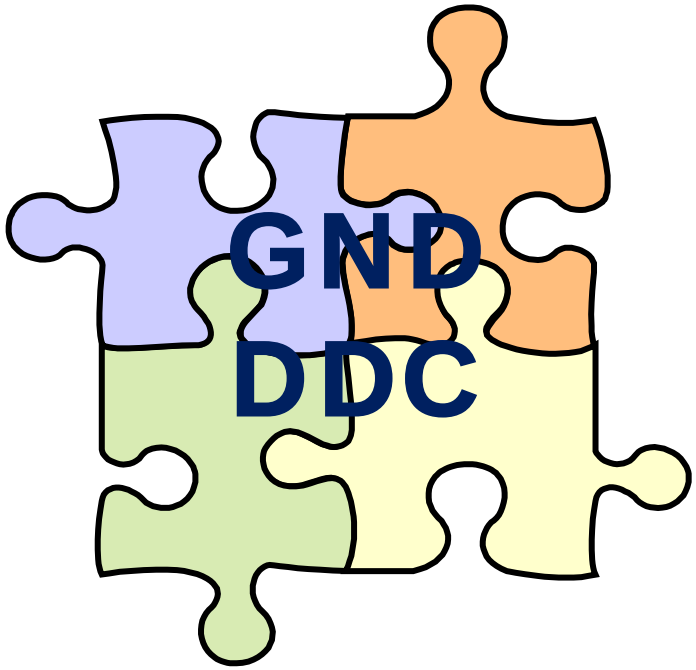
Automated subject cataloging processes to become standard procedures

Reuse of subject metadata to be increased

Maintenance of the GND to be strengthened

} Stra-tegic focus

Advancement of subject cataloging concepts



Automatic subject cataloging

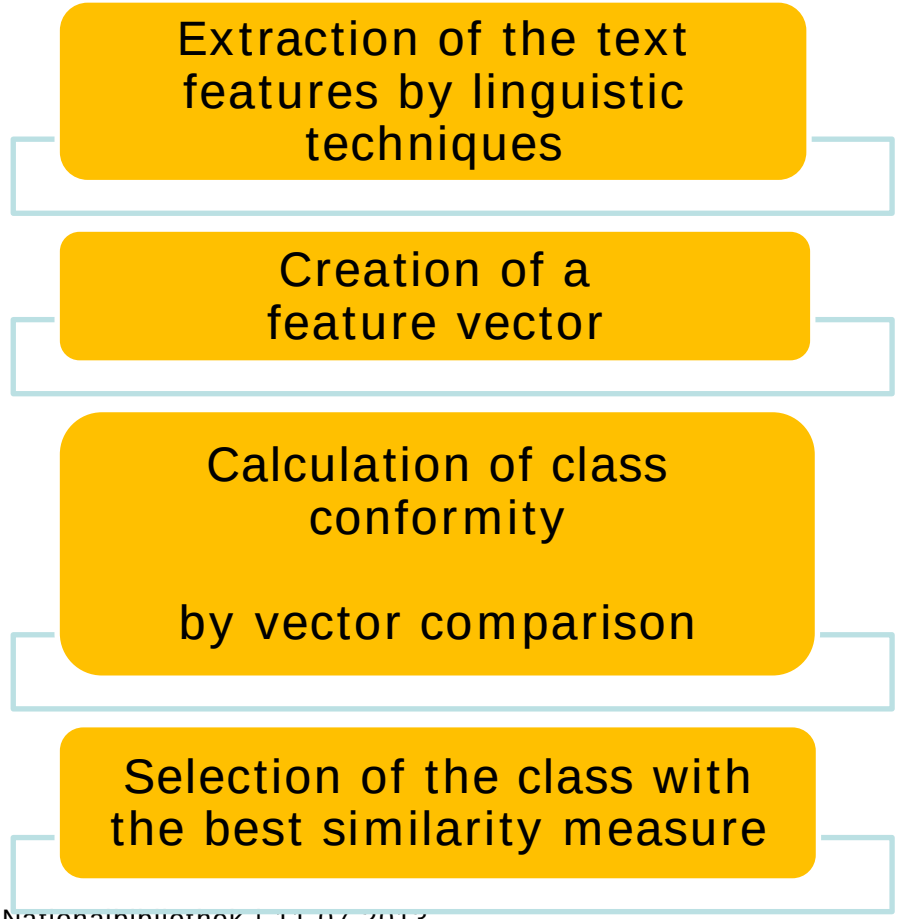
- Petrus project since 2009
- Subject cataloging
  - Automated assignment of DDC subject groups
  - Automated indexing (assignment of subject headings)
- Cooperation with averbis GmbH Freiburg
- Focus on online publications

Automatic subject cataloging
Automatic assignment of DDC subject groups

Machine-learning procedures:

Mathematical models are created by machine learning methods using manually catalogued publications

In production since 2012: processing of German and English online publications



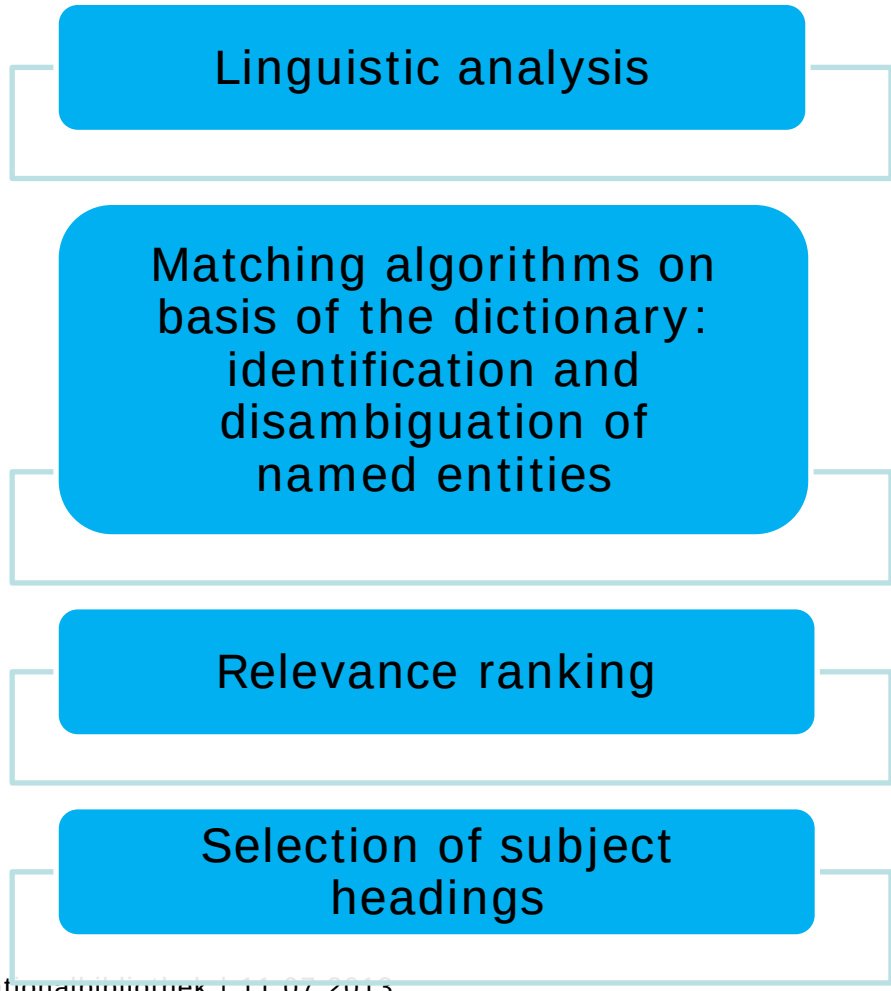
Automatic subject cataloging
Subject indexing

Basis: Dictionaries with the controlled vocabulary of GND: currently

- ca. 330.000 persons
- ca. 172.000 topical terms
- ca. 158.000 geographical terms

In process of startup in 2013: processing of German online dissertations

Step by step integration of more types of publications



Automatic subject cataloging
Use of GND

- Automatic subject indexing will naturally
  - not be in conformity with the RSWK
  - nor will the GND be processed in exactly the same way.
- Despite all possible improvements: a difference remains.

- **The main target is to <u>improve the retrieval</u>.**
- Misleading search terms are like poison: the dose makes the difference!

Specific maintenance of the controlled vocabulary is the crucial point to achieve a sufficient indexing quality!

Automatic subject cataloging
Subject indexing

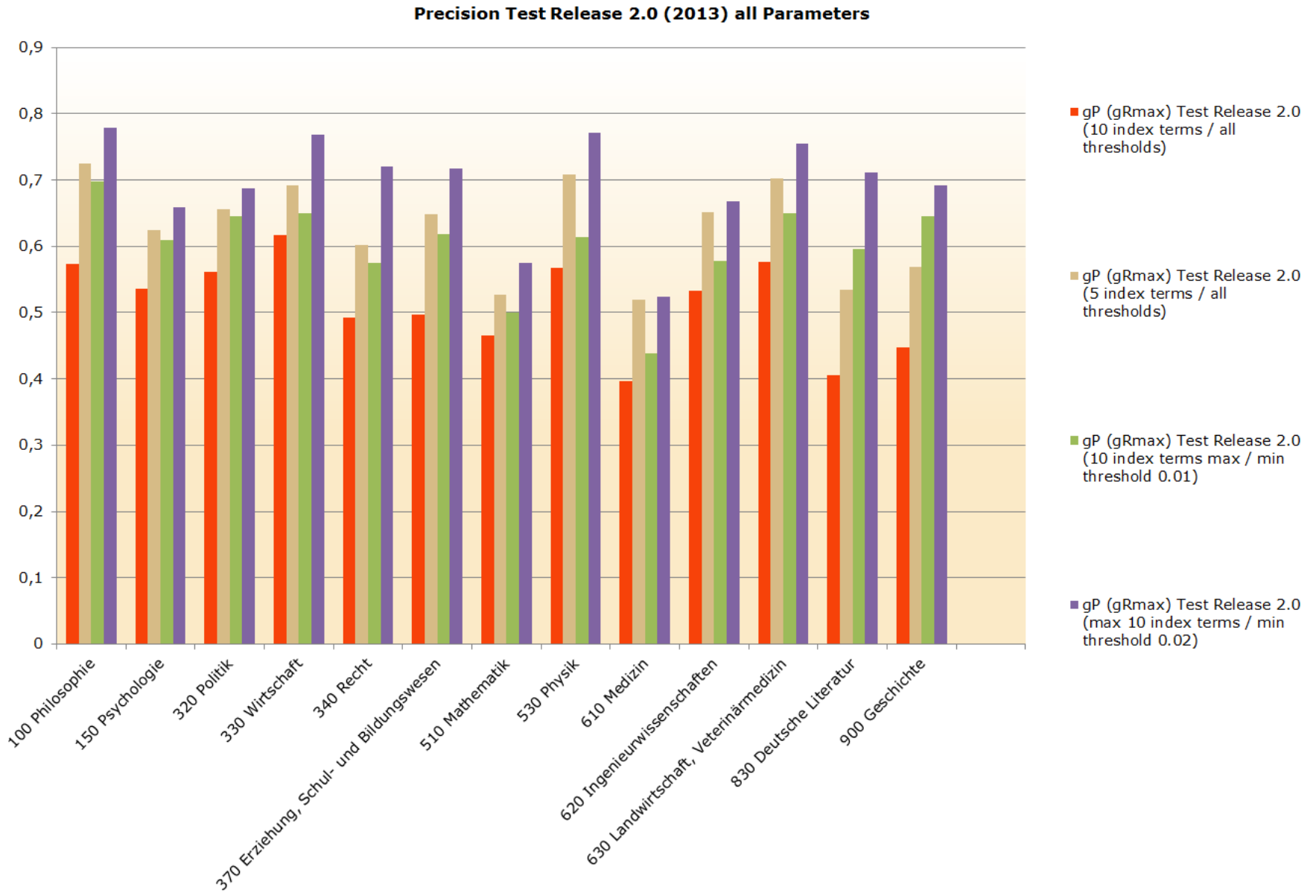
DEUTSCHE NATIONAL BIBLIOTHEK

Example: http://d-nb.info/972736786

- Title: Ehrenschutz von Soldaten in Deutschland und anderen Staaten
- Subject group: 340 legislation

SW	IDN	KW	Bewertung
|s|Ehrenschutz	041390172	0.838	sehr nützlich
|s|Soldat	040554090	0.027	sehr nützlich
|s|Ehre	041307674	0.018	nützlich
|g|Deutschland	040118827	0.011	sehr nützlich
|s|Mörder	041703499	0.007	nützlich
|s|Rechtsgut	040487997	0.006	falsch
|s|Verfassung	04062787X	0.006	wenig nützlich
|s|Beleidigung	041209915	0.005	nützlich
|s|Gesetz	040206602	0.004	wenig nützlich
|s|Rechtsprechung	041157109	0.004	wenig nützlich

- Evaluation: +
- Missing terms: |s|Rechtsvergleich


Precision Test Release 2.0 (2013) all Parameters
0,9
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
100 Philosophie
150 Psychologie
320 Politik
330 Wirtschaft
340 Recht
370 Erziehung, Schul- und Bildungswesen
510 Mathematik
530 Physik
610 Medizin
620 Ingenieurwissenschaften
630 Landwirtschaft, Veterinärmedizin
830 Deutsche Literatur
900 Geschichte
gP (gRmax) Test Release 2.0 (10 index terms / all thresholds)
gP (gRmax) Test Release 2.0 (5 index terms / all thresholds)
gP (gRmax) Test Release 2.0 (10 index terms max / min threshold 0.01)
gP (gRmax) Test Release 2.0 (max 10 index terms / min threshold 0.02)

Automatic subject cataloging
Principles of quality control

- No intellectual post-processing of automatically generated data
- Work with samples
  - To assess systematic errors
  - To gain more data for training purposes
- Quality statistics to identify trends

- Transparency in terms of data provenance and quality when recording and delivering data
- Gradual improvement of quality of subject cataloging through optimizing/combining and supplementing methods

Automatic subject cataloging
Data about data:
provenance and quality labels

0500 Oaf

0600 ro;ra;pb

3000 |m|!1031805168!Colverson, Michael

4000 Bist Du schon wach oder schläfst Du noch? : In Geiselhaft der Großbanken und Großkonzerne werden wir entweder geschoren oder geschlachtet! / Michael Colverson

5050 330**EiD**2013-03-05

5050 330**EpD**2013-03-08

5050 330**EmH**dnb**$K**0,8**$D**2013-03-01

5050 000**EaH**onx**$D**2013-03-01

Order illustrates the ranking:

$Ei	– intellectual provenance
$Ep	– taken from parallel edition
$Em	– automatic classification
$Ea	– external supplier

New MARC field 883: Machine-generated Metadata Provenance

Data reuse
Exchange of subject data between records

For ~ 1/3 of the online publications (about 200.000 so far) subject data (subject headings, DDC notations, DDC subject groups) are transferred from parallel editions

In process of planning:

- Identification of other editions of the same work independently from the type of material (i.e. other manifestations or expressions in the sense of FRBR/RDA)
- Identification of other publications of the same author

Data reuse

Subject headings from ZBW Kiel

- Retrospective enrichment of records of Series B with subject headings of ZBW Kiel
- Mapping of headings of Standardthesaurus Wirtschaft with GND headings
- Match&Merge-Procedure to identify identical titles in DNB and ZBW (processed so far: 547.000 titles)
- ca. 41.000 titles enrichted, 1976-2010
- **In planning:** regular reuse of ZBW data

0500 Aa
0600 rb;sf
1100 2010
...
3000 !124733956!Hurd, Michael D.
3001 !135811317!Rooij, Maarten**$c**van
3002 !121853063!Winter, Joachim
4000 Stock market expectations of Dutch households / Michael Hurd ; Maarten van Rooij ; Joachim Winter. MEA, Mannheim Research Institute for the Economics of Aging ; [Universität Mannheim, Fakultät für Volkswirtschaftslehre und Statistik]
...
5050 330
5550 [ckw]!040422038!Niederlande
5550 [ckw]!95774482X!Kapitalertrag
5550 [ckw]!04214003X!Anlageverhalten
5550 [ckw]!945908504!Kapitalmarktforschung
5550 [ckw]!040237443!Haushalt
5550 [ckw]Geschichte 2004-2006
5560 [stw]2004-2006
5560 [stw](DE-STW)122855*Kapitalertrag
5560 [stw](DE-STW)186746*Anlageverhalten
5560 [stw](DE-STW)133981*Privater Haushalt
5560 [stw](DE-STW)174335*Niederlande
[0101] leipzig dnb <101a>

...

Data reuse
Other projects

- Reuse of genre terms for fiction and childrens‘ books provided by the marketing agency of the German book publishing and sellers‘ trade
- Project:Can verbal access points be harvested from CrissCross links?

Concluding remarks

- Develop a concept integrating various subject data generated in different ways:
  For which publications which subject data are created?
- RDA and FRBR as a chance: subject data on the work level!
- Subject cataloging rules should take new methods of creating subject access into account
- Shift from manually processing individual publications to „mass procedures“, quality control and data management needs new staff skills and a change management

Thank you!

Questions?

Contact: u.junger@dnb.de

www.dnb.de