DEUTSCHE NATIONAL BIBLIOTHEK

Frank Busse

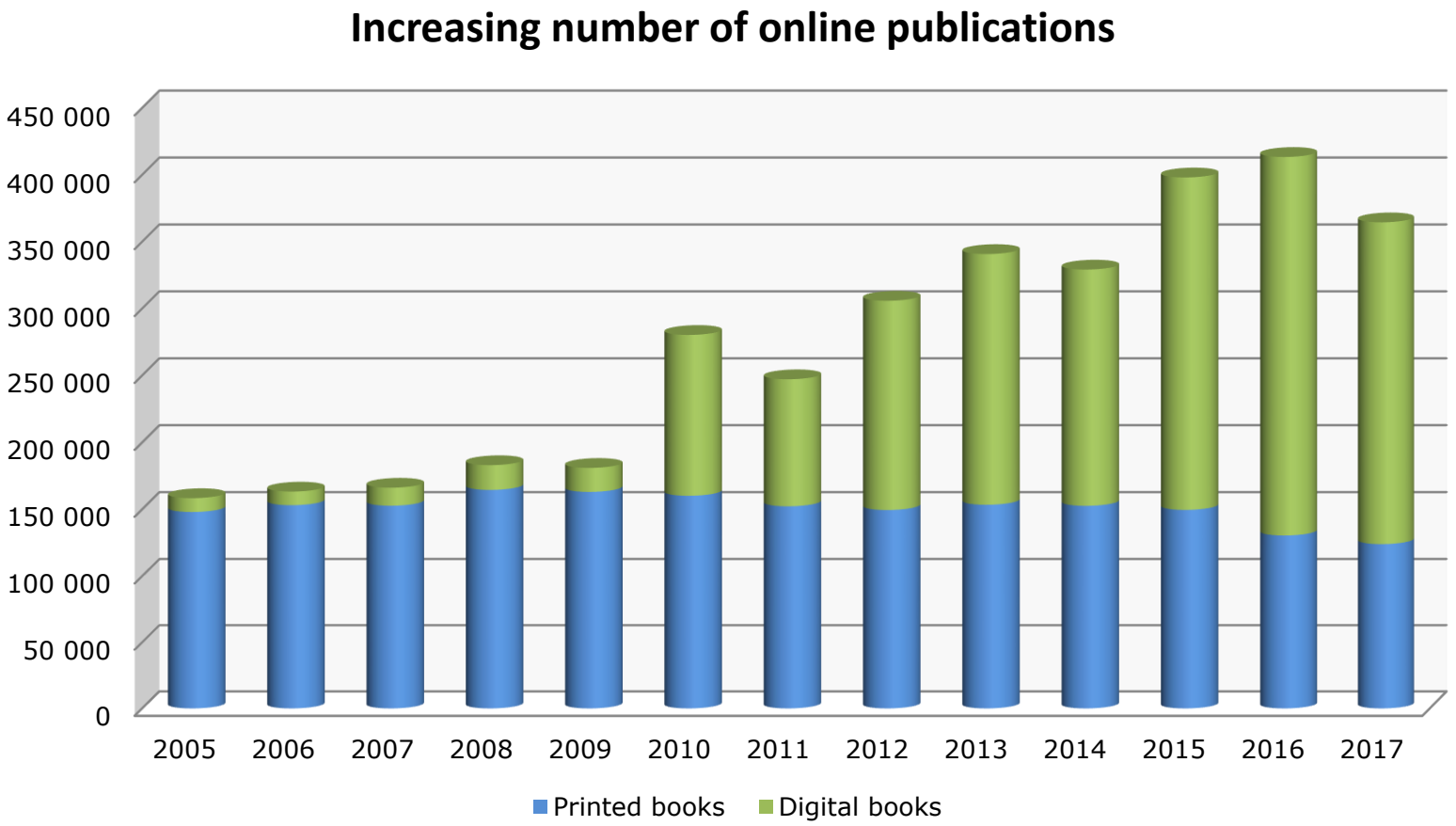
Automatic classification at the DNB

Outline

1. **General Information**
2. **Technical background**
3. **Field of use**
   - DDC Subject Categories
   - DDC Short Numbers
4. **Results & Problems**
5. **Outlook**

General Information

Automated Cataloguing – why?



Timeline

- 2009 Start of PETRUS project
- 2010 Ceasing of intellectual cataloguing of online publications
- 2012 Automatic classification / DDC Subject Categories
- 2014 Automatic indexing
- 2015 Automatic classification / DDC Short Numbers
- 2015 PETRUS project completed

Technical background

Machine Learning

- Support Vector Machine (SVM)
- Supervised learning (Learning by example)
- Pattern recognition
- Classifying unknown objects

Software

- Averbis GmbH / Freiburg im Breisgau
- Averbis Extraction Platform (AEP)
- Future improvements

Workflow

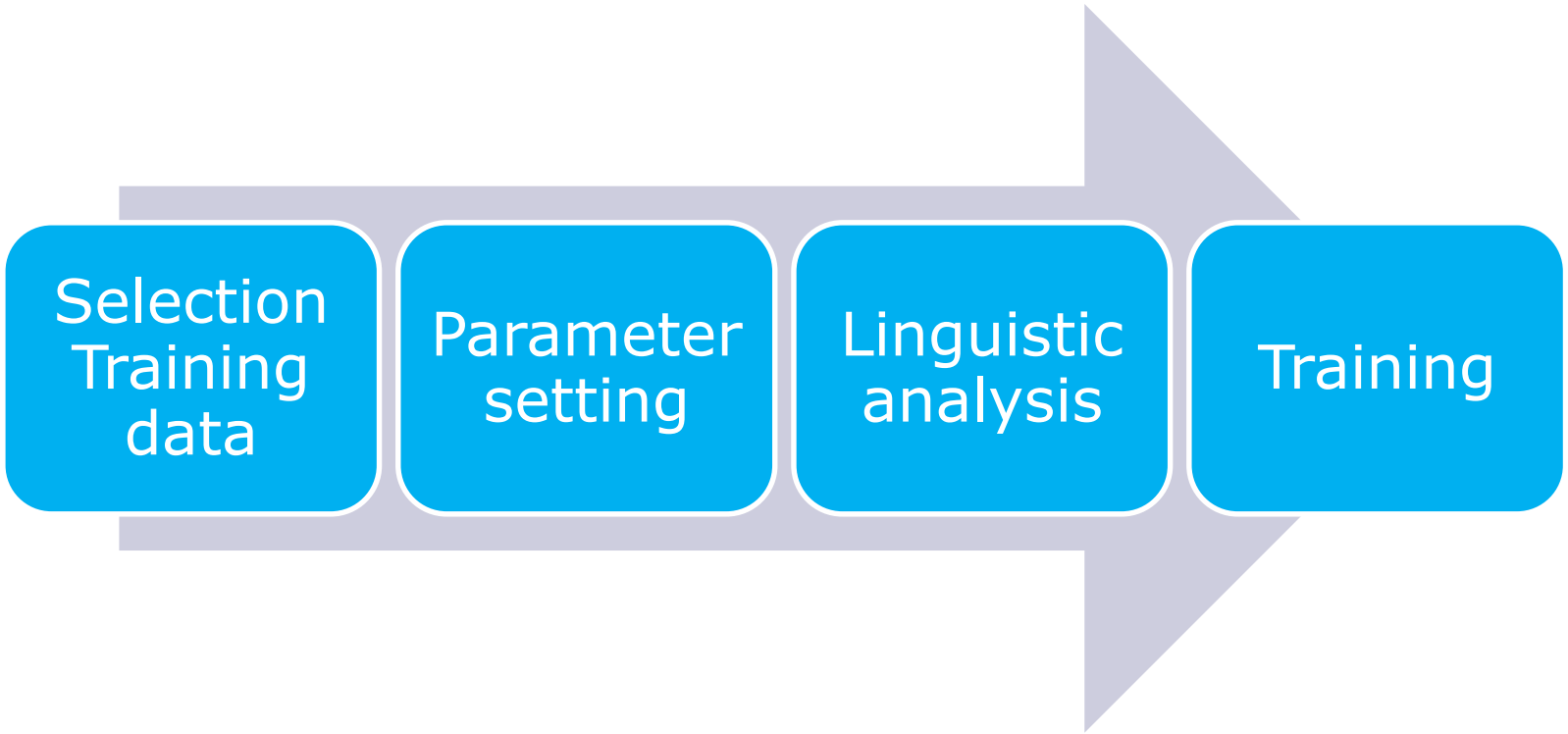
Training

- Base
- Create a model
- Software:
  - Averbis Software

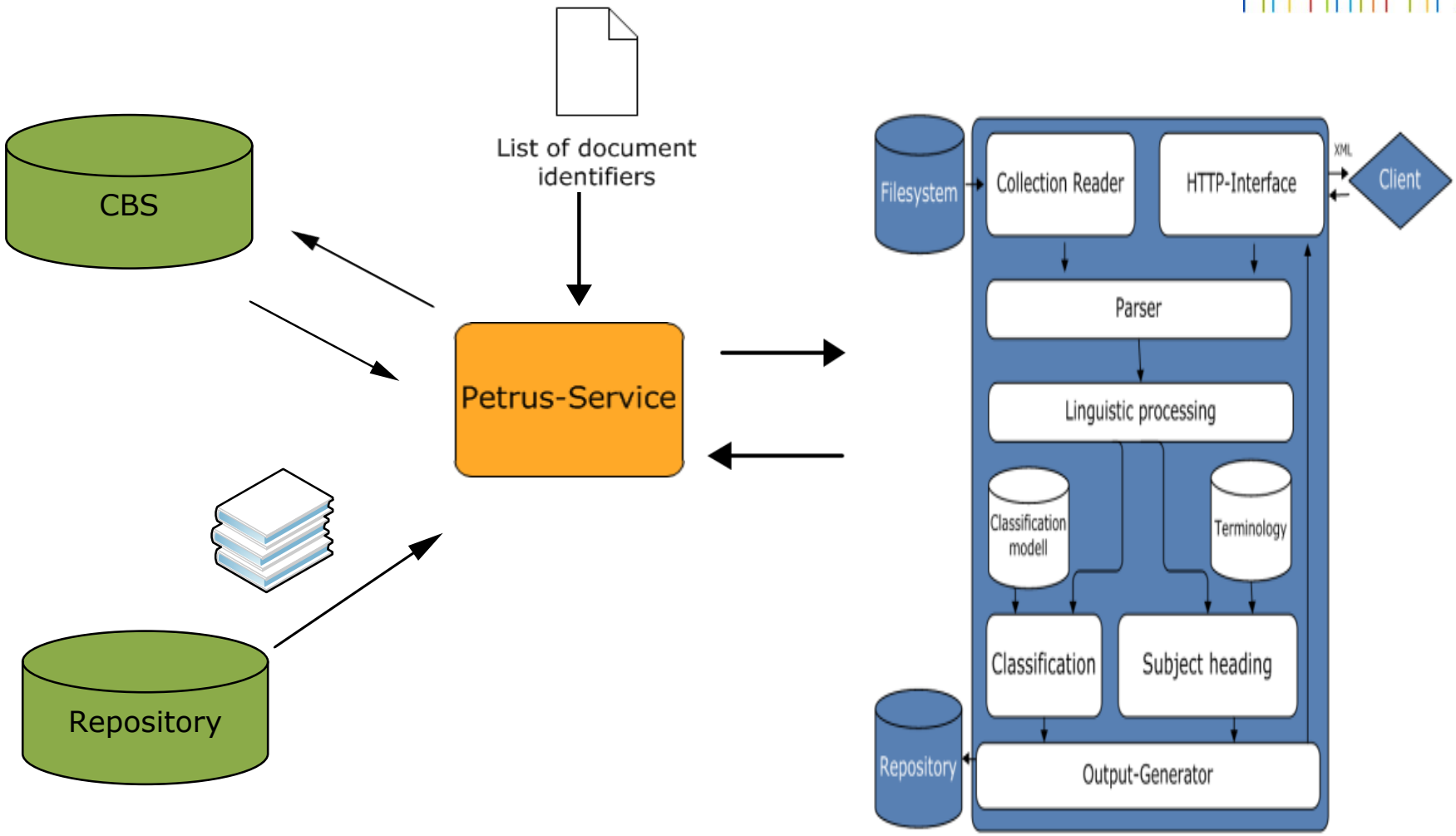
Routine

- Daily processing of new online publications
- Retro-processing
- Software:
  - Averbis Software
  - DNB Interface
  - CBS

Training


Selection Training data
Parameter setting
Linguistic analysis
Training

Routine


List of document identifiers
CBS
Petrus-Service
Repository
Filesystem
Collection Reader
HTTP-Interface
XML
Client
Parser
Linguistic processing
Classification modell
Terminology
Classification
Subject heading
Repository
Output-Generator

Field of use:

DDC Subject Categories

DDC Subject Categories

- Since 2004
- Based on Dewey Decimal Classification (DDC)
- 102 categories

Automatic Classification

- Start: 2012
- Method: machine learning / SVM
- Document type:
  - All online publications / without fiction
  - PDF (since 2012)
  - Epub (since 2015)
  - Language Ger/Eng
- Volume: 1.362.719 objects (04/2018)

Field of use:

DDC Short Numbers

What is a DDC Short Number?

DDC-SC	610
DDC	618.92398009435123090511
Short Number	618.92 ~~398~~009~~435123090511~~

DNB-SC	004
DDC	005.82
Short Number	005.8 ~~2~~

DDC-SC	300
DDC	303.625088297095 6
Short Number	303.6 ~~25~~088~~2970956~~

DDC-SC	610
DDC	616.4624061071
Short Number	616.4 ~~624~~061~~071~~

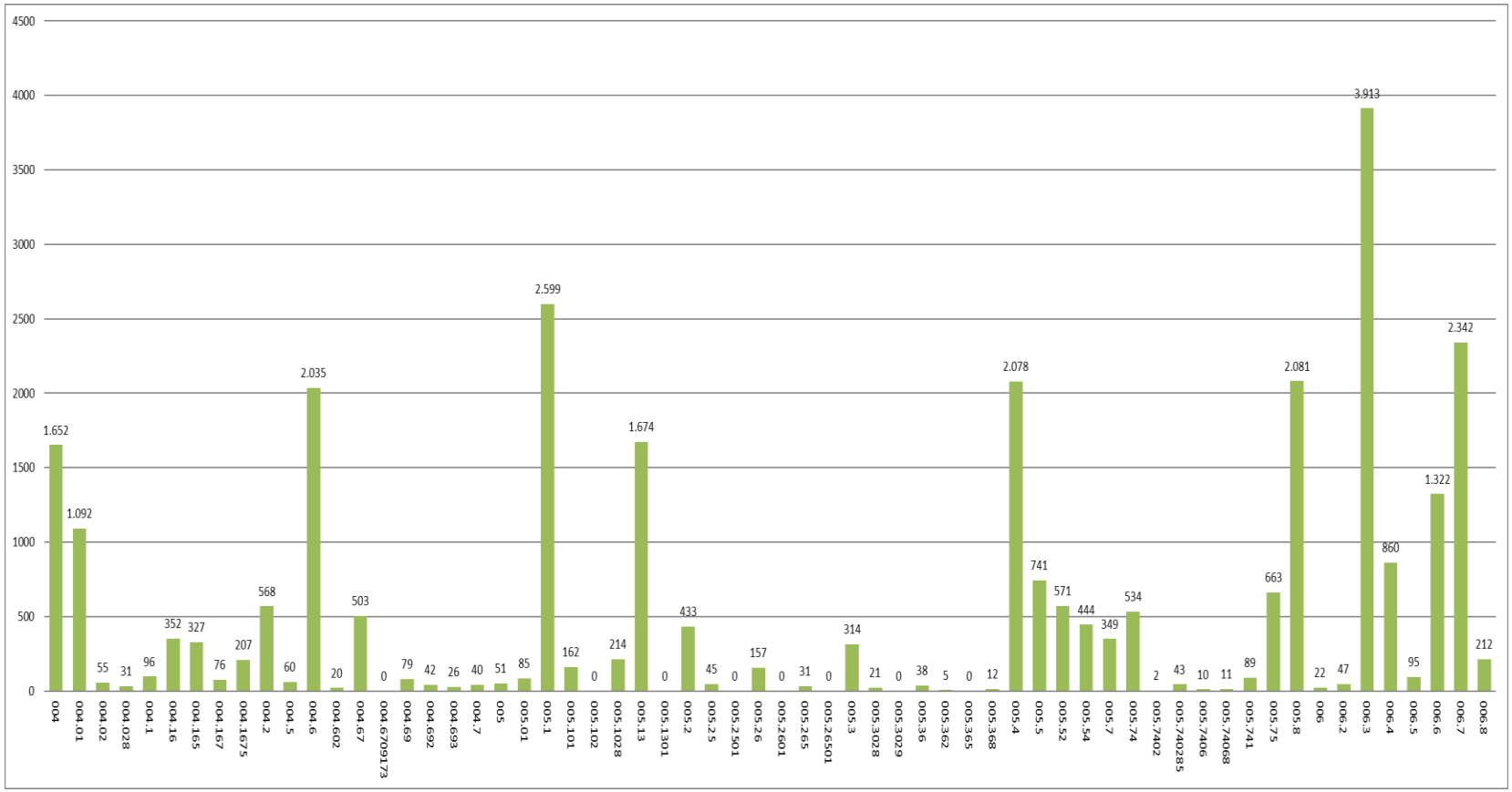
How to create a DDC Short Number system

- Step I : Starting point: DDC Abridged Edition 15

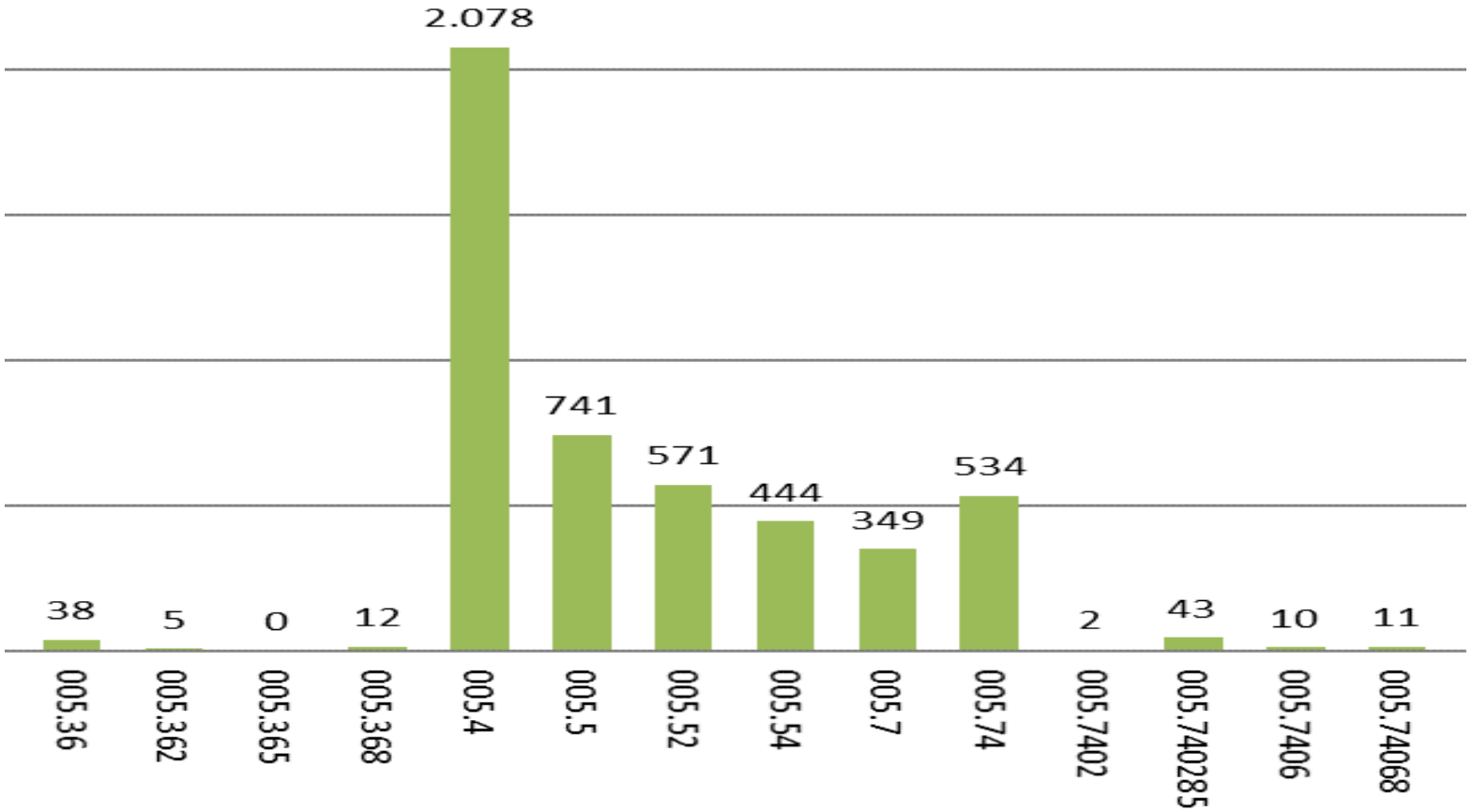
How to create a DDC Short Number system

- Step I : Starting point: DDC Abridged Edition 15
- Step II : Data analysis

Step II: Data analysis



Step II: Data analysis close view



How to create a DDC Short Number system

- Step I : Starting point: DDC Abridged Edition 15
- Step II : Data analysis
- Step III : Customize
  - Throw away unneeded numbers
  - Identify required numbers

How to create a DDC Short Number System

- Step I : Starting point: DDC Abridged Edition 15
- Step II : Data analysis
- Step III : Customize
  - Throw away unneeded numbers
  - Identify required numbers
- Step IV : Testing
  - Test-analyse-adjust, test-analyse-adjust, …

How to create a DDC Short Number System

- Step I : Starting point: DDC Abridged Edition 15
- Step II : Data analysis
- Step III : Customize
  - Throw away unneeded numbers
  - Identify required numbers
- Step IV : Testing
  - Test-analyse-adjust, test-analyse-adjust, …
- Step V : Put into operation

Progress overview (04/2018)

- **Operational Short Numbers**
  - 004 Computer science
  - 300 Social sciences, sociology, anthropology
  - 540 Chemistry
  - 610 Medicine and health
- **Short Numbers in preparation**
  - 370 Education
  - 130 Parapsychology, occultism
  - 710 Landscaping and area planning
  - 720 Architecture
  - ….

Identification

- Identifier (label) for all Short Numbers
- Identifier (label) for all machine-determined short numbers
- Data exchange in MARC 21 will start in May 2018

Identification / DNB Portal

Link zu diesem Datensatz	http://d-nb.info/1127024027
Titel	Insomnia: Medical Sleep Disorder & Diagnosis / Md Belal Bin Heyat
Person(en)	Heyat, Md Belal Bin (Verfasser)
Ausgabe	1. Auflage
Verlag	Hamburg : Anchor Academic Publishing
Zeitliche Einordnung	Erscheinungsdatum: 2017
Umfang/Format	Online-Ressourcen, 56 Seiten (pdf)
Andere Ausgabe(n)	Elektronische Reproduktion: ISBN: 9783960675891
Persistent Identifier	URN: urn:nbn:de:101:1-2017030745
URL	http://www.anchor-publishing.com/e-book/337729/insomnia-medical-sleep-disorder-diagnosis (Verlag)
ISBN/Einband/Preis	978-3-96067-089-6
EAN	9783960670896
Sprache(n)	Englisch (eng)
Anmerkungen	Lizenzpflichtig. - Vom Verlag als Druckwerk on demand und/oder als E-Book angeboten Langzeitarchivierung gewährleistet
DDC-Notation	616.8 (maschinell ermittelte DDC-Kurznotation)
Sachgruppe(n)	610 Medizin, Gesundheit

Identification / DNB Portal

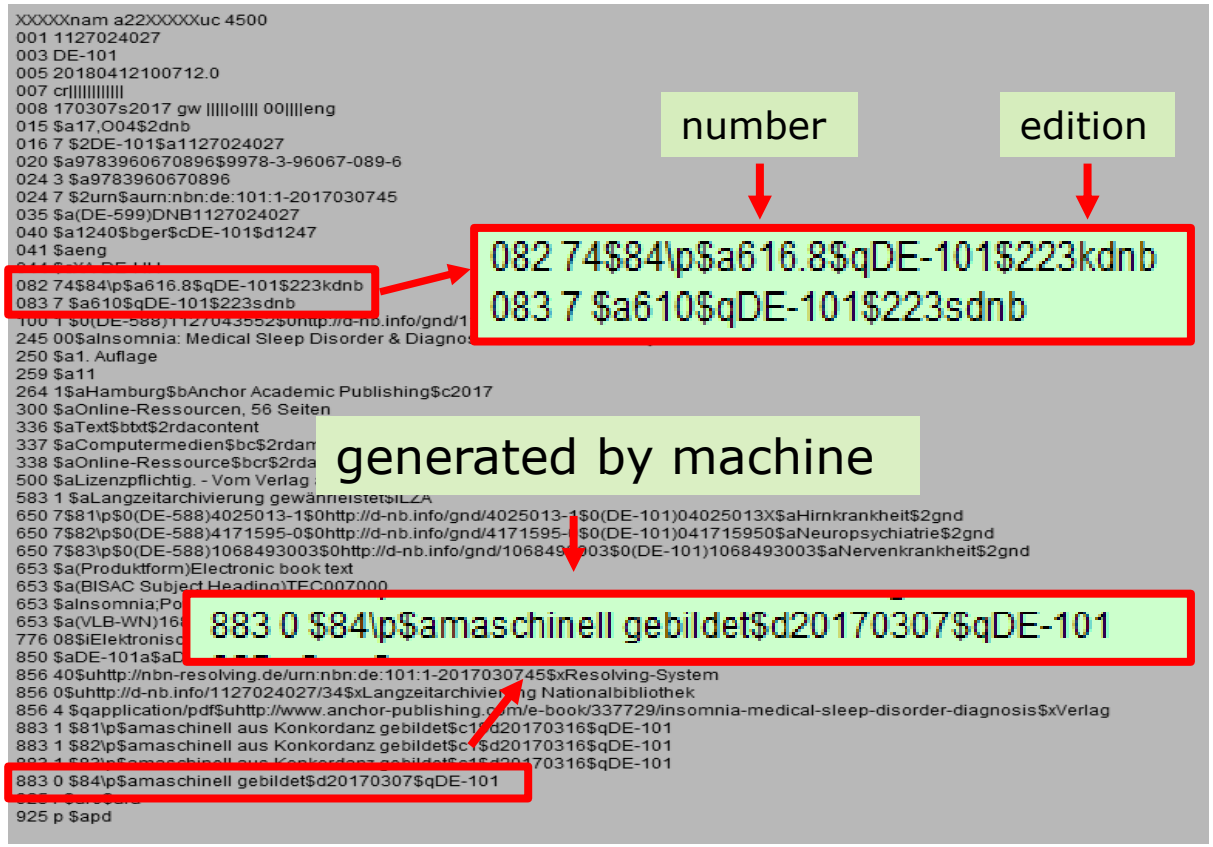
Link zu diesem Datensatz	http://d-nb.info/1127024027
Titel	Insomnia: Medical Sleep Disorder & Diagnosis / Md Belal
Person(en)	Heyat, Md Belal Bin (Verfasser)
Ausgabe	1. Auflage
Verlag	Hamburg : Anchor Academic Publishing
Ze	
Ur	
An	
Pe	
UR	
ISBN/Einband/Preis	978-3-96067-089-6
EAN	9783960670896
Sprache(n)	Englisch (eng)
Anmerkungen	Lizenzpflichtig. - Vom Verlag als Druckwerk on demand und/oder als E-Book angeboten Langzeitarchivierung gewährleistet
DDC-Notation	616.8 (maschinell ermittelte DDC-Kurznotation)
Sachgruppe(n)	610 Medizin, Gesundheit

generated by machine

Identification / MARC 21

```
XXXXXnam a22XXXXXuc 4500
001 1127024027
003 DE-101
005 20180412100712.0
007 cr||||||||||||
008 170307s2017 gw |||||o|||| 00||||eng
015 $a17,O04$2dnb
016 7 $2DE-101$a1127024027
020 $a9783960670896$9978-3-96067-089-6
024 3 $a9783960670896
024 7 $2urn$aurn:nbn:de:101:1-2017030745
035 $a(DE-599)DNB1127024027
040 $a1240$bger$cDE-101$d1247
041 $aeng
044 $cXA-DE-HH
082 74$84\p$a616.8$qDE-101$223kdnb
083 7 $a610$qDE-101$223sdnb
100 1 $0(DE-588)1127043552$0http://d-nb.info/gnd/1127043552$0(DE-101)1127043552$aHeyat, Md Belal Bin$eVerfasser$4aut
245 00$aInsomnia: Medical Sleep Disorder & Diagnosis$cMd Belal Bin Heyat
250 $a1. Auflage
259 $a11
264 1$aHamburg$bAnchor Academic Publishing$c2017
300 $aOnline-Ressourcen, 56 Seiten
336 $aText$btxt$2rdacontent
337 $aComputermedien$bc$2rdamedia
338 $aOnline-Ressource$bcr$2rdacarrier
500 $aLizenzpflichtig. - Vom Verlag als Druckwerk on demand und/oder als E-Book angeboten
583 1 $aLangzeitarchivierung gewährleistet$iLZA
650 7$81\p$0(DE-588)4025013-1$0http://d-nb.info/gnd/4025013-1$0(DE-101)04025013X$aHirnkrankheit$2gnd
650 7$82\p$0(DE-588)4171595-0$0http://d-nb.info/gnd/4171595-0$0(DE-101)041715950$aNeuropsychiatrie$2gnd
650 7$83\p$0(DE-588)1068493003$0http://d-nb.info/gnd/1068493003$0(DE-101)1068493003$aNervenkrankheit$2gnd
653 $a(Produktform)Electronic book text
653 $a(BISAC Subject Heading)TEC007000
653 $aInsomnia;Power Spectral Density;Diagnosis;Sleep Disorder;Short Time Frequency;EEG Signal
653 $a(VLB-WN)1684
776 08$iElektronische Reproduktion$z9783960675891
850 $aDE-101a$aDE-101b
856 40$uhttp://nbn-resolving.de/urn:nbn:de:101:1-2017030745$xResolving-System
856 0$uhttp://d-nb.info/1127024027/34$xLangzeitarchivierung Nationalbibliothek
856 4 $qapplication/pdf$uhttp://www.anchor-publishing.com/e-book/337729/insomnia-medical-sleep-disorder-diagnosis$xVerlag
883 1 $81\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 1 $82\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 1 $83\p$amaschinell aus Konkordanz gebildet$c1$d20170316$qDE-101
883 0 $84\p$amaschinell gebildet$d20170307$qDE-101
925 r $aro$ara
925 p $apd
```

Identification / MARC 21



Results & Problems

Results 2017:
DDC Subject Categories

Classified objects: 126.580

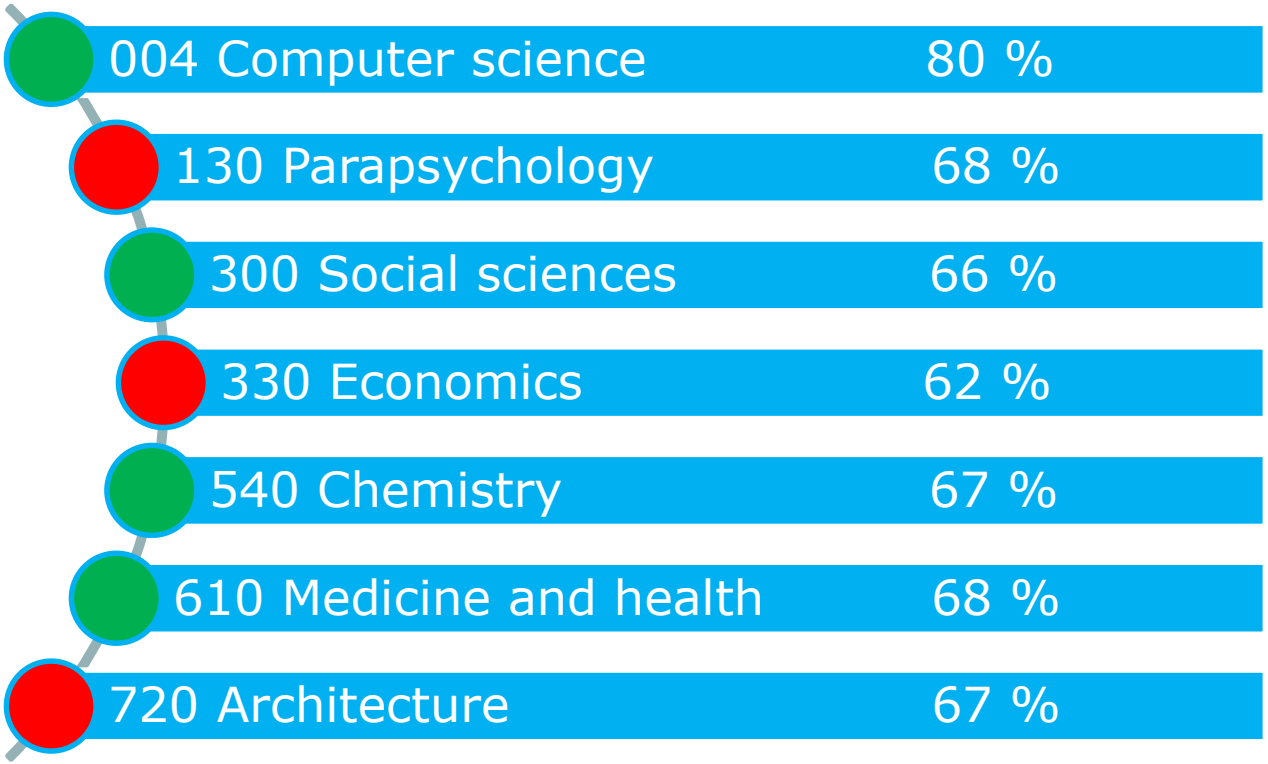
Sample check: 16.321 (13%)

Result: 77 % correct

2017 Tops & Flops

Category	Total 2017	Sample 2017	Match
340 Law	5.308	9.82	91 %
610 Medicine and health	16.296	1.435	90 %
650 Management	7.004	1.541	87 %
…			
960 History of Africa	23	7	29 %
980 History of S. America	10	5	20 %
000 Generalities	202	55	20 %

Results: DDC Short Numbers (Testing)



Results: DDC Short Numbers

T-330_A2-02

Micro_F1	0,614		Macro_F1	0,5695		Mapping	V6		xval	5
Korpora	DE/EN		Objekte	26.915		Text	40.000			
Titel GW	10		004N$a GW	0		004K$a GW	0			
330	561	**561**	0,73							
		330.01	281	**281**	0,52					
				330.0151	551	**551**	0,80			
			330.071	149	**149**	0,73				
		330.09	161	**161**	0,49					
	330.1	288	**288**	0,48						
		330.12	631	**631**	0,64					
		330.15	222	**222**	0,50					
	330.9	230	**230**	0,39						
		330.94	89	**89**	0,51					
			330.943	170	**170**	0,23				
331	160	**160**	0,50							
		331.01	157	**157**	0,63					
	331.1	21	**21**	0,31						
		331.11	271	**271**	0,53					
		331.12	385	**385**	0,56					
					331.12042	192	**192**	0,66		
			331.127	171	**171**	0,62				
		331.13	174	**174**	0,49					
	331.2	319	**319**	0,53						

Results 2017:
Short Numbers for 610 Medicine and health

Classified objects: 20.123

Sample check: 1.567 (8%)

Result: 69 % correct

Problems

- Consequential errors:
  - the notation depends on the assigned subject category
- Many small subject categories
- (No) Quality standards
- How to handle changes in the DDC

Outlook

Future challenges

- Improve results
- Adjustment and reorganization of processes
- Creating sets of Short Numbers for all DDC Subject Categories

Thank you for your attention!

Questions?

Frank Busse
German National Library
Section Automatic Indexing, Online Publications

f.busse@dnb.de