DEUTSCHE
NATIONAL
BIBLIOTHEK

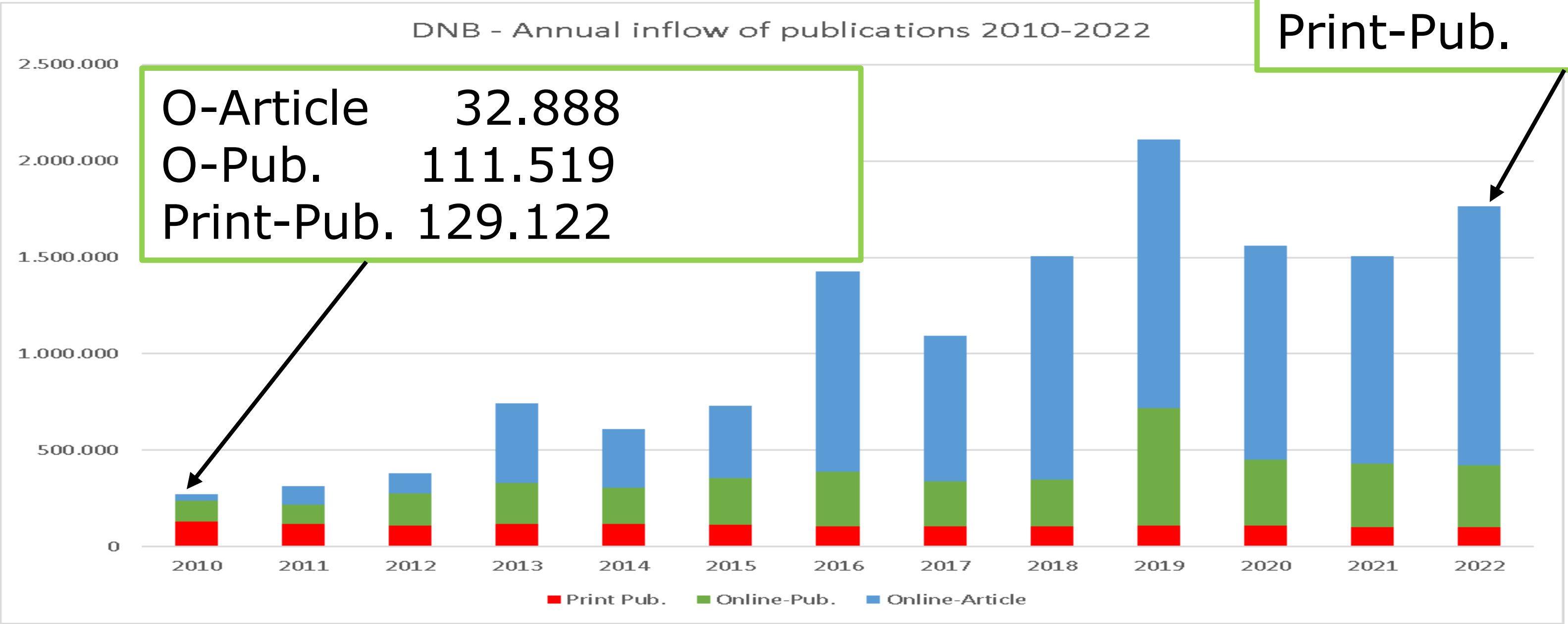
Frank Busse

10 years of automated DDC classification at DNB

Outline

1. **General Information / History**
2. **Use Case**
3. **Current results**
4. **Outlook**

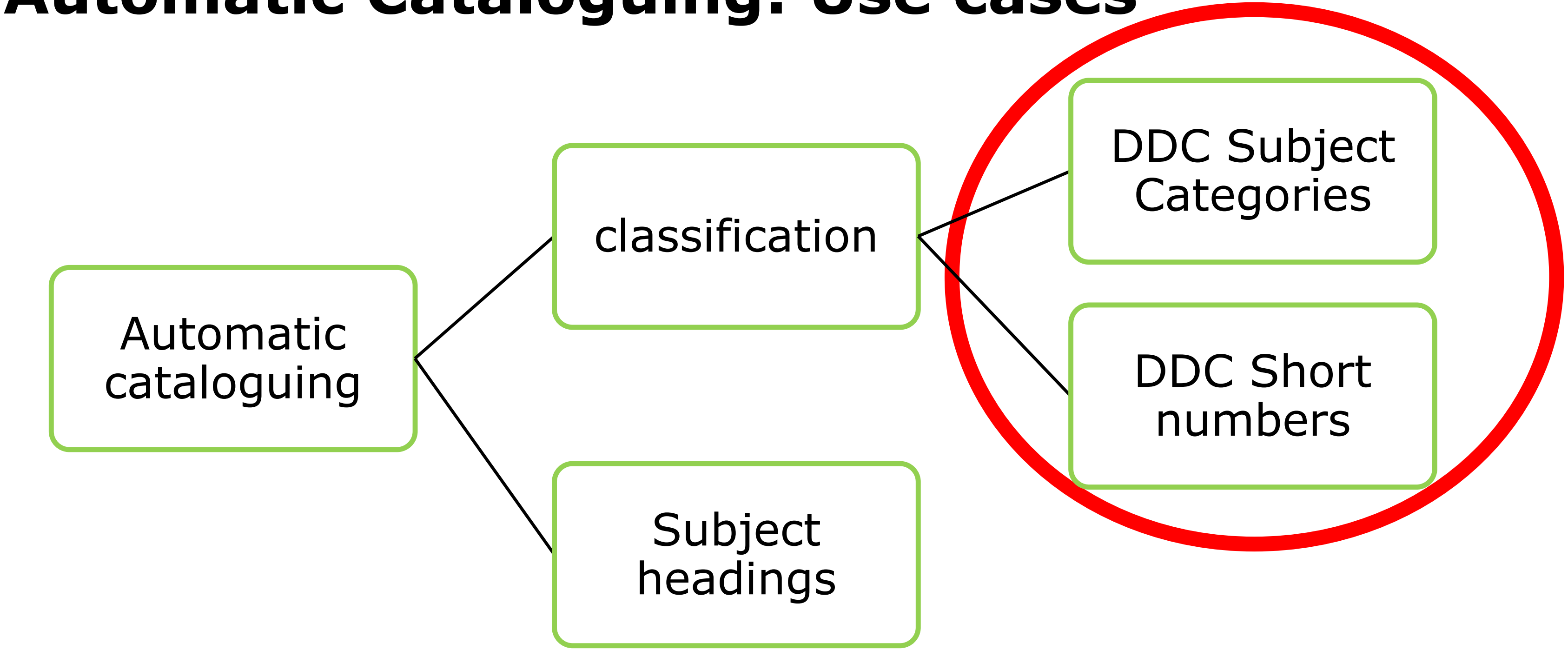
Automatic Cataloguing – why?



O-Article	32.888
O-Pub.	111.519
Print-Pub.	129.122

O-Article	1.342.525
O-Pub.	323.210
Print-Pub.	99.242

Automatic Cataloguing: Use cases



Automatic cataloguing

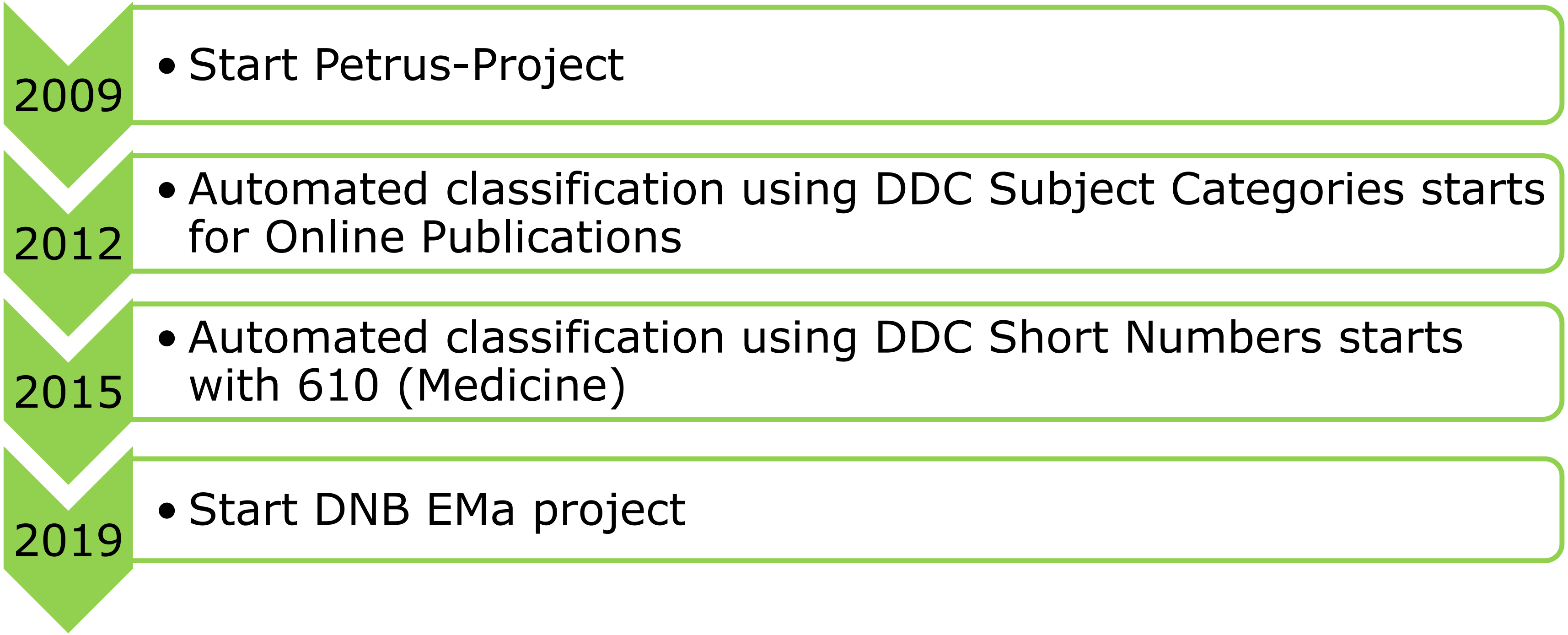
classification

Subject headings

DDC Subject Categories

DDC Short numbers

Milestones in automatic Cataloguing



2009

- Start Petrus-Project

2012

- Automated classification using DDC Subject Categories starts for Online Publications

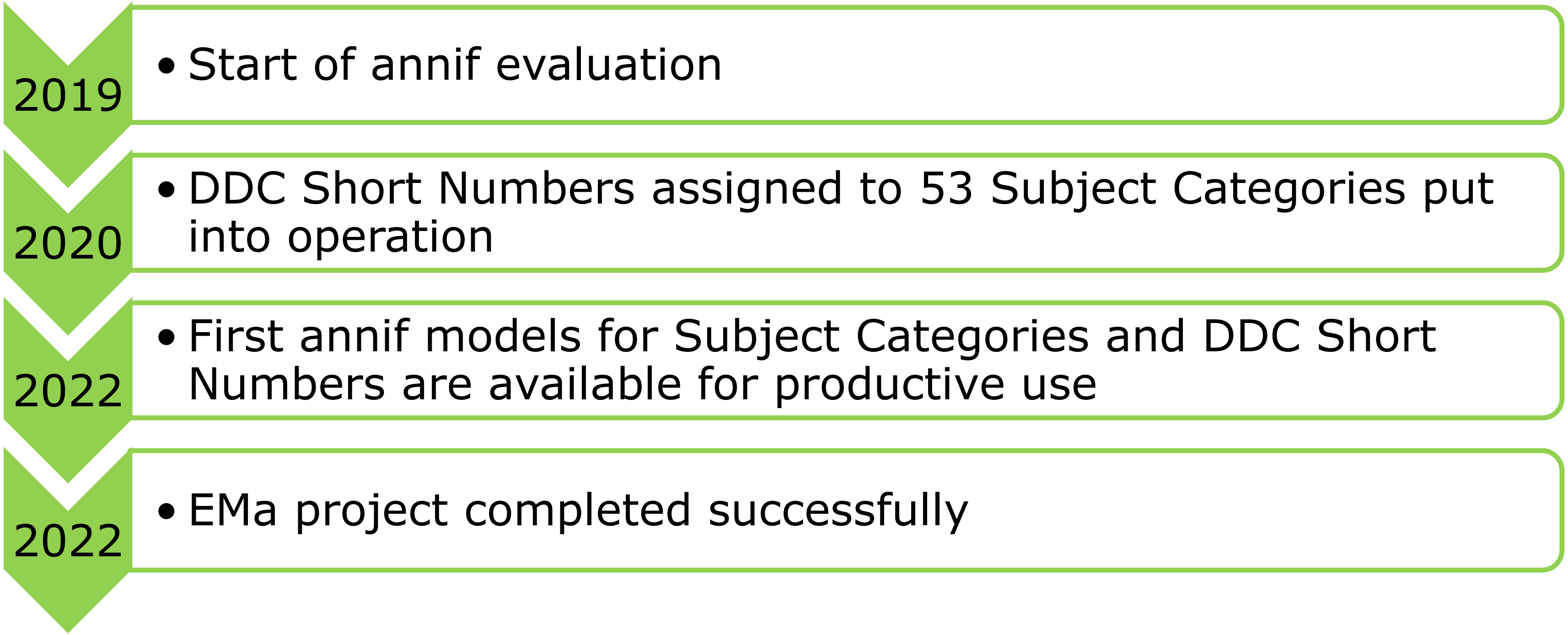
2015

- Automated classification using DDC Short Numbers starts with 610 (Medicine)

2019

- Start DNB EMa project

Milestones in automatic Cataloguing



2019
- Start of annif evaluation

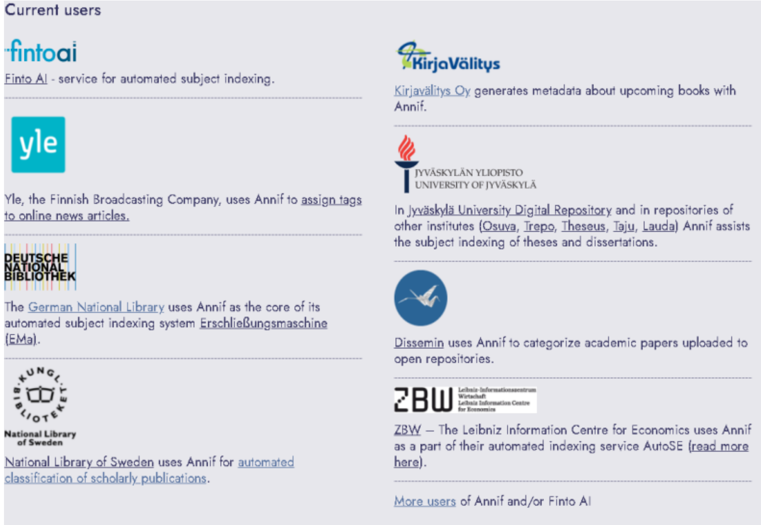
2020
- DDC Short Numbers assigned to 53 Subject Categories put into operation

2022
- First annif models for Subject Categories and DDC Short Numbers are available for productive use

2022
- EMa project completed successfully

annif

- Open source toolbox developed at the National Library of Finland
- Uses different tools for natural language processing & machine learning
- Works multilingual
- Vocabulary in SKOS or simple TSV
- Open source
- Growing international community



Current users

Finto AI - service for automated subject indexing.

Yle, the Finnish Broadcasting Company, uses Annif to assign tags to online news articles.

The German National Library uses Annif as the core of its automated subject indexing system Erschließungsmaschine (EMa).

National Library of Sweden uses Annif for automated classification of scholarly publications.

Kirjavälitys Oy generates metadata about upcoming books with Annif.

In Jyväskylä University Digital Repository and in repositories of other institutes (Osuva, Trepo, Theseus, Taju, Lauda) Annif assists the subject indexing of theses and dissertations.

Dissemin uses Annif to categorize academic papers uploaded to open repositories.

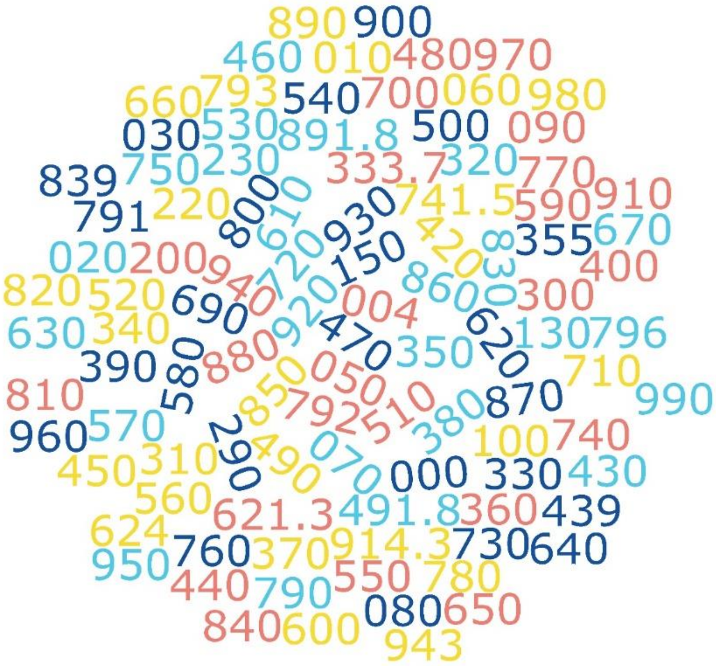
ZBW – The Leibniz Information Centre for Economics uses Annif as a part of their automated indexing service AutoSE (read more here).

More users of Annif and/or Finto AI

fintoai
yle
DEUTSCHE NATIONAL BIBLIOTHEK
KUNGL. BIBLIOTEKET
National Library of Sweden
KirjaVälitys
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ
ZBW
Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

DDC Subject Categories

103 Subject categories



5,221,530 Objects classified by machine

most frequently assigned category is 610 with 1.5 million objcets

DEUTSCHE
NATIONAL
BIBLIOTHEK

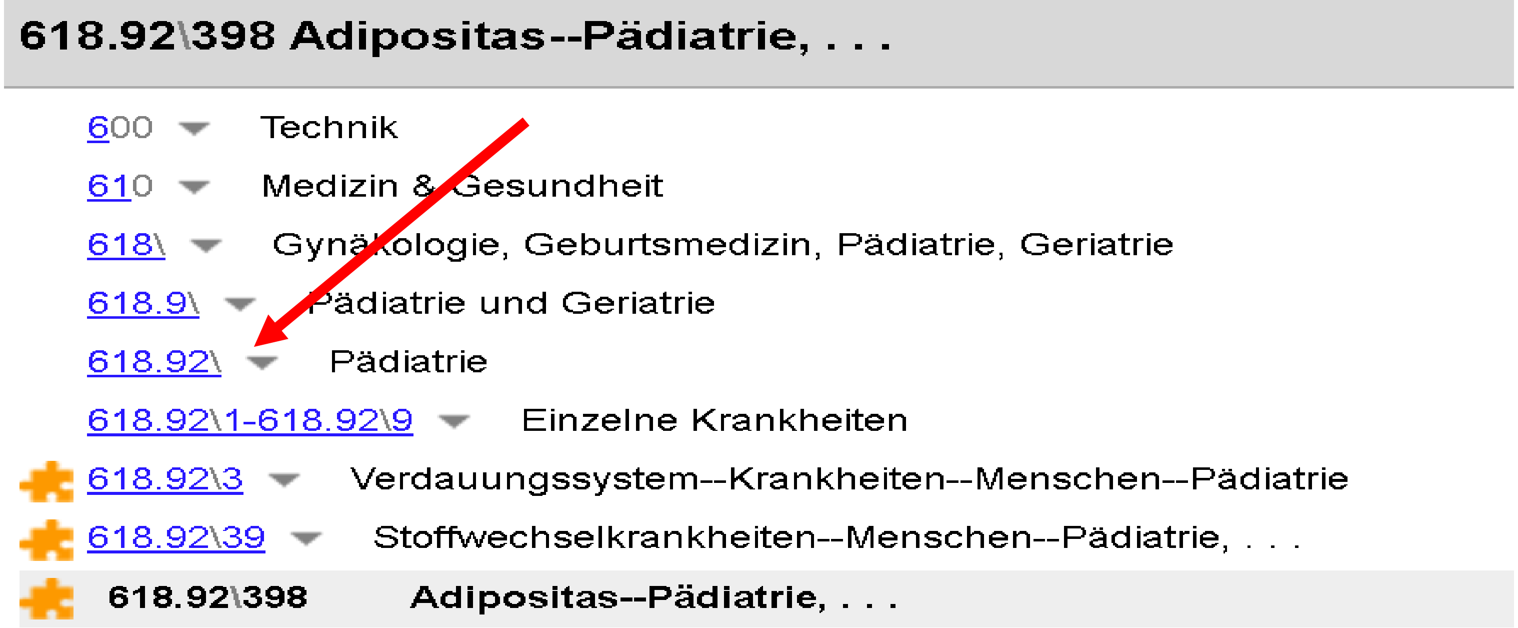
DDC Short Numbers

- DDC Short Numbers are part of the entire DDC system
- DDC Short Numbers are a professionally selected and fixed set of DDC notations
- The starting point for the development of the DDC short notations is the DDC short edition (Abridged DDC, edition 15)
- Their selection takes into account the volume of literature in the DNB
- The notation length and thus the degree of specification therefore often deviates from the short edition

DDC Short Numbers Examples

DDC	618.9239800943	Pediatrics--Obesity--Germany
DDC Subject Category	610	Medicine, Health
DDC Short Number	618.92	Pediatrics
DDC	658.404	Project management
DDC Subject Category	650	Management
DDC Short Number	658.404	Project management

DDC Short Numbers WebDewey



618.92\398 Adipositas--Pädiatrie, . . .

600	Technik
610	Medizin & Gesundheit
618\	Gynäkologie, Geburtsmedizin, Pädiatrie, Geriatrie
618.9\	Pädiatrie und Geriatrie
618.92\	Pädiatrie
618.92\1-618.92\9	Einzelne Krankheiten
618.92\3	Verdauungssystem--Krankheiten--Menschen--Pädiatrie
618.92\39	Stoffwechselkrankheiten--Menschen--Pädiatrie, . . .
618.92\398	**Adipositas--Pädiatrie, . . .**

DDC Short Numbers

53 Subject
categories

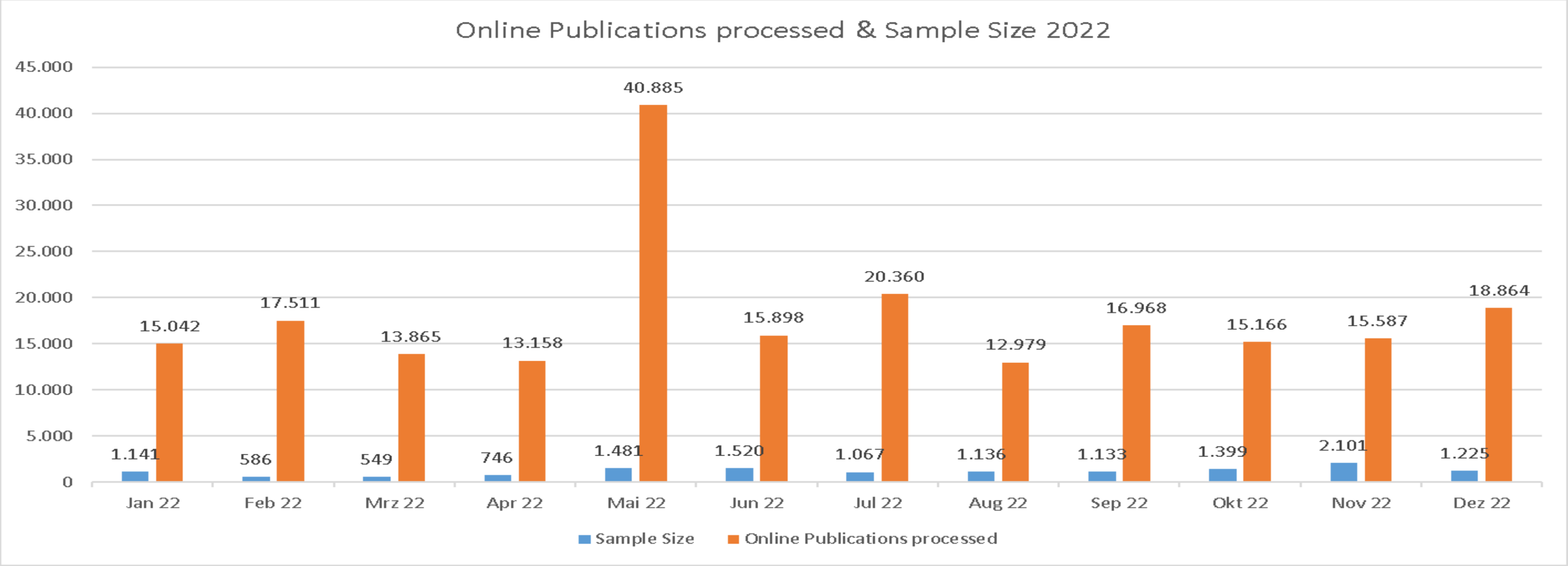
2.9 Mio.
Objects classified

600
490
740
791
620
790
390
510
400
624
491.8
630
570
300
700
590
290
610
670
500
520
540
793
430
460
621.3
690
550
640
220
230
560
710
650
792
760
660
530
130
750
439
200
004
150
420
720
380
580
480
450
730
440
470

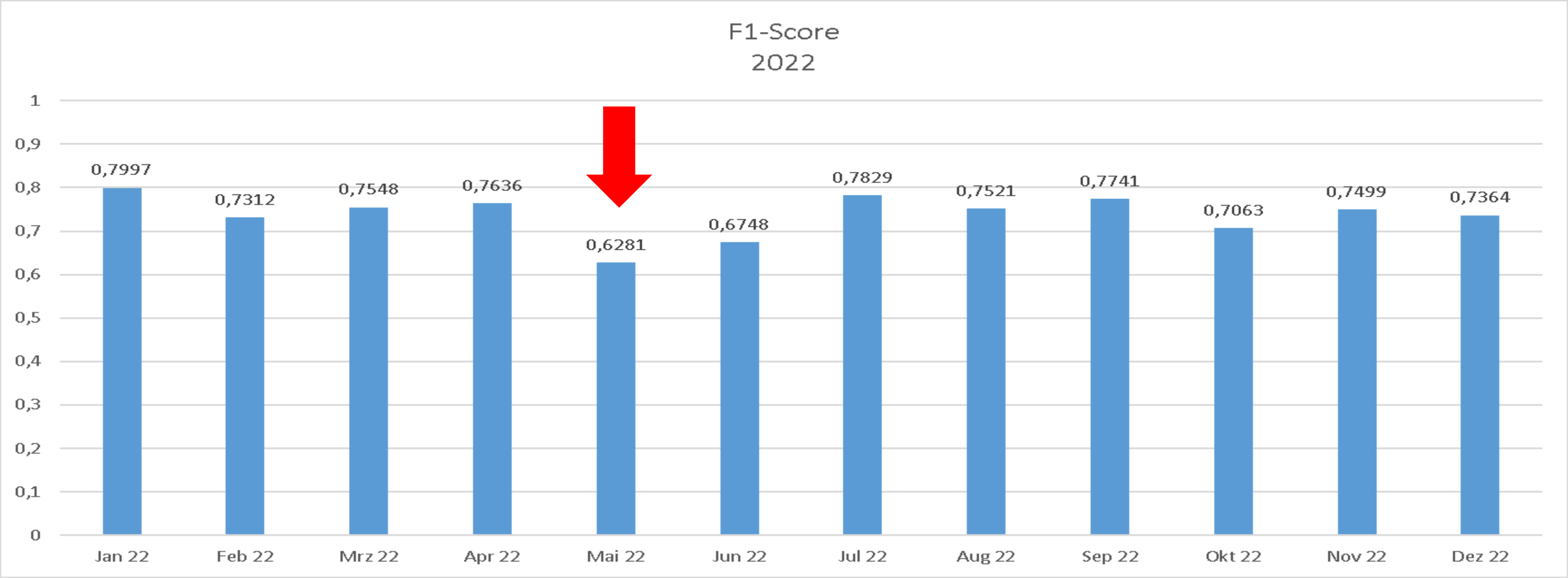
2,105 Short Numbers

1.5 Mio. DDC Short Numbers come from the category Medicine

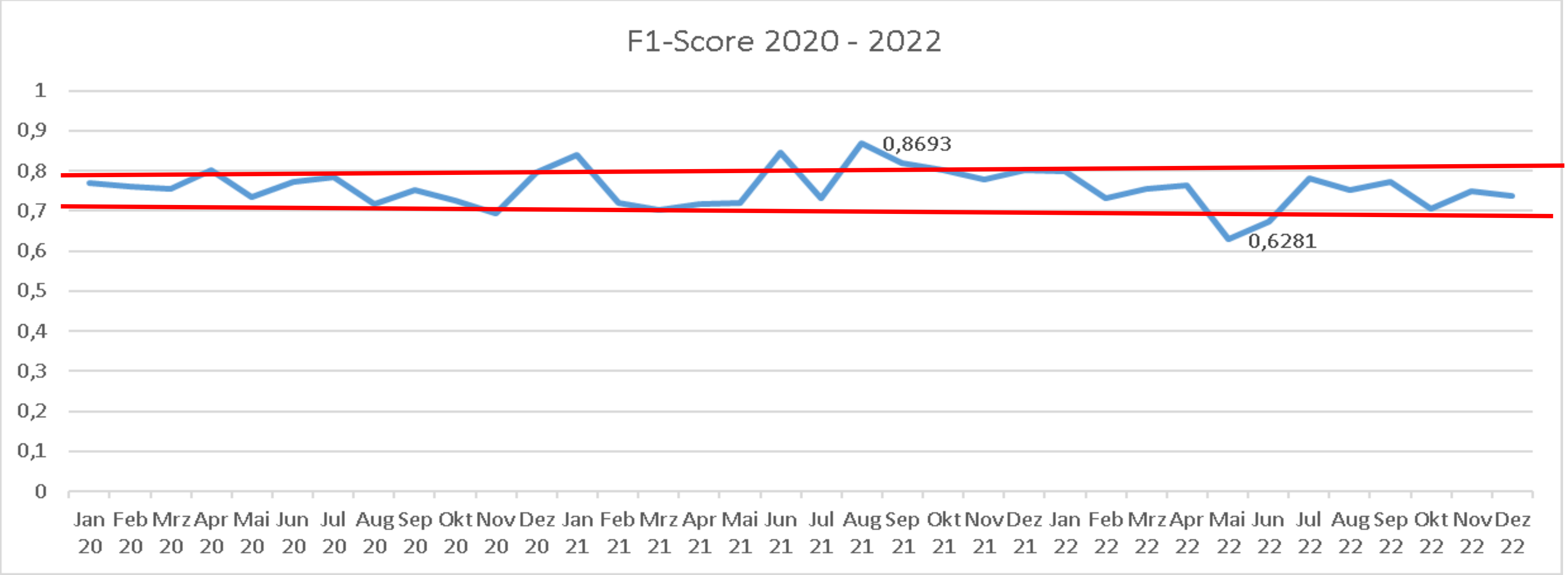
Sample size 2022



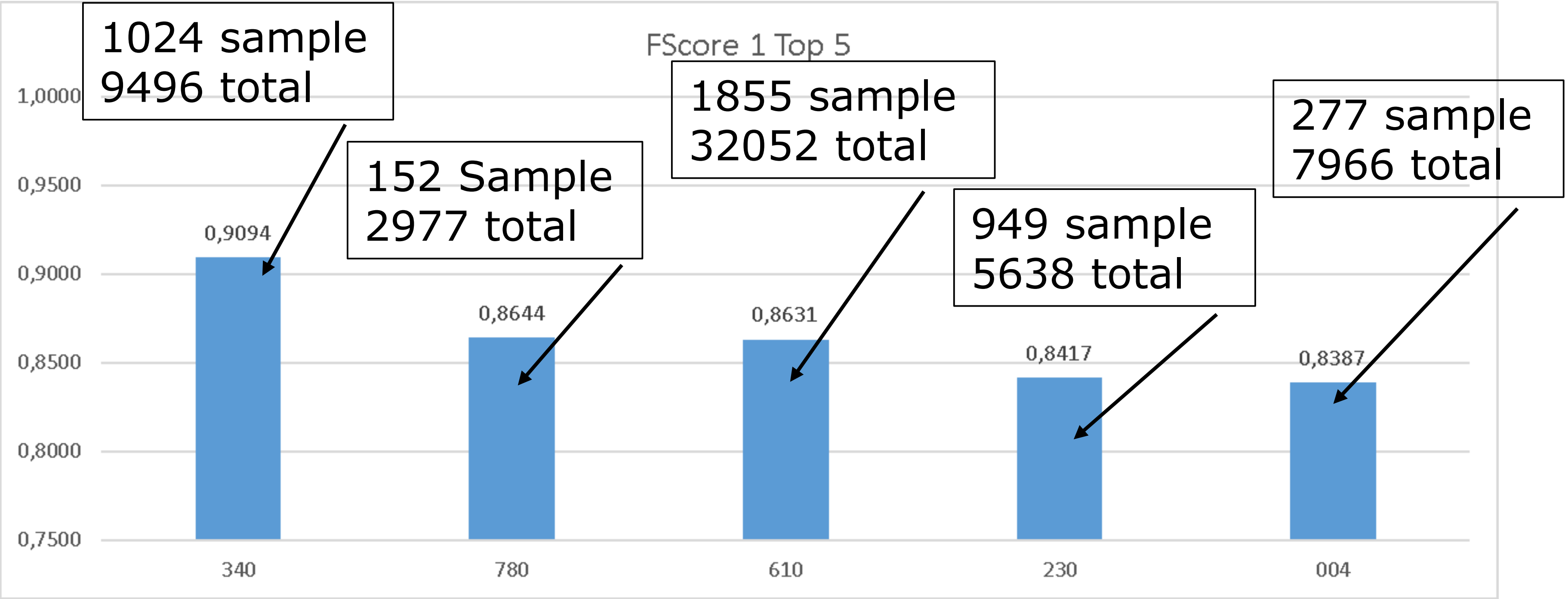
Results DDC Subject Categories 2022



Results DDC Subject Categories 2020-2022



Highest F1-Score Subject Categories 2022



1024 sample
9496 total

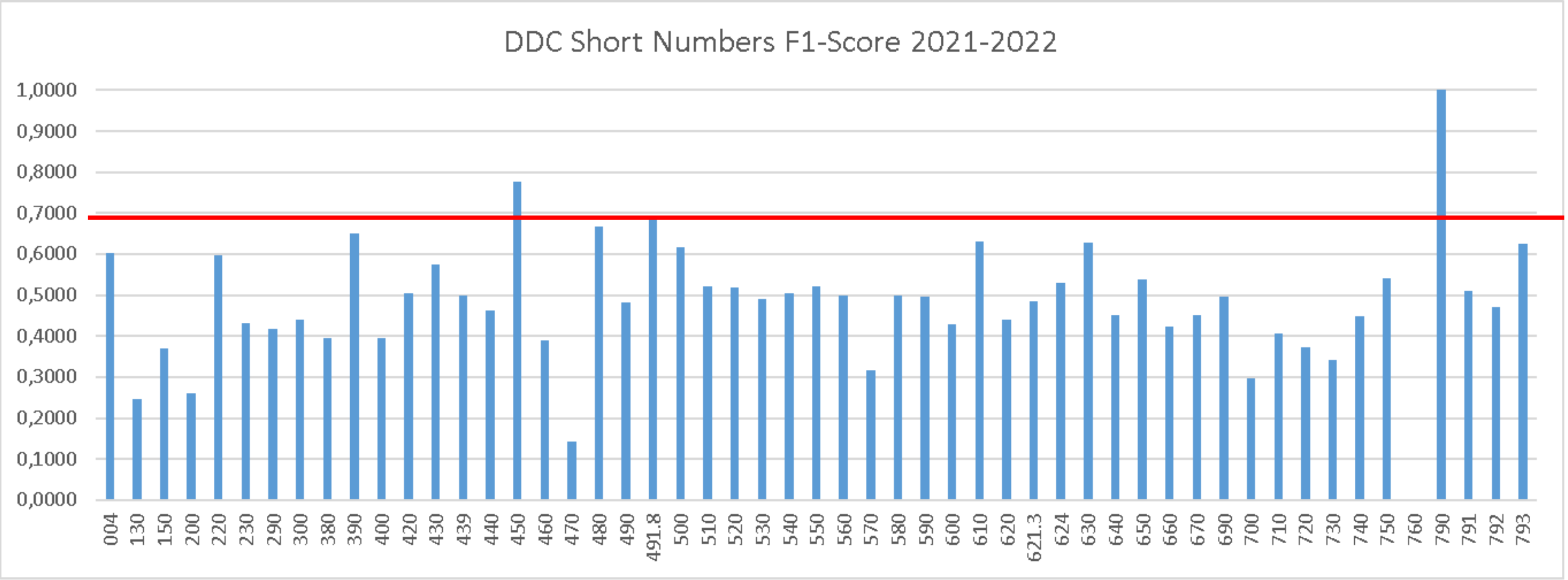
152 Sample
2977 total

1855 sample
32052 total

949 sample
5638 total

277 sample
7966 total

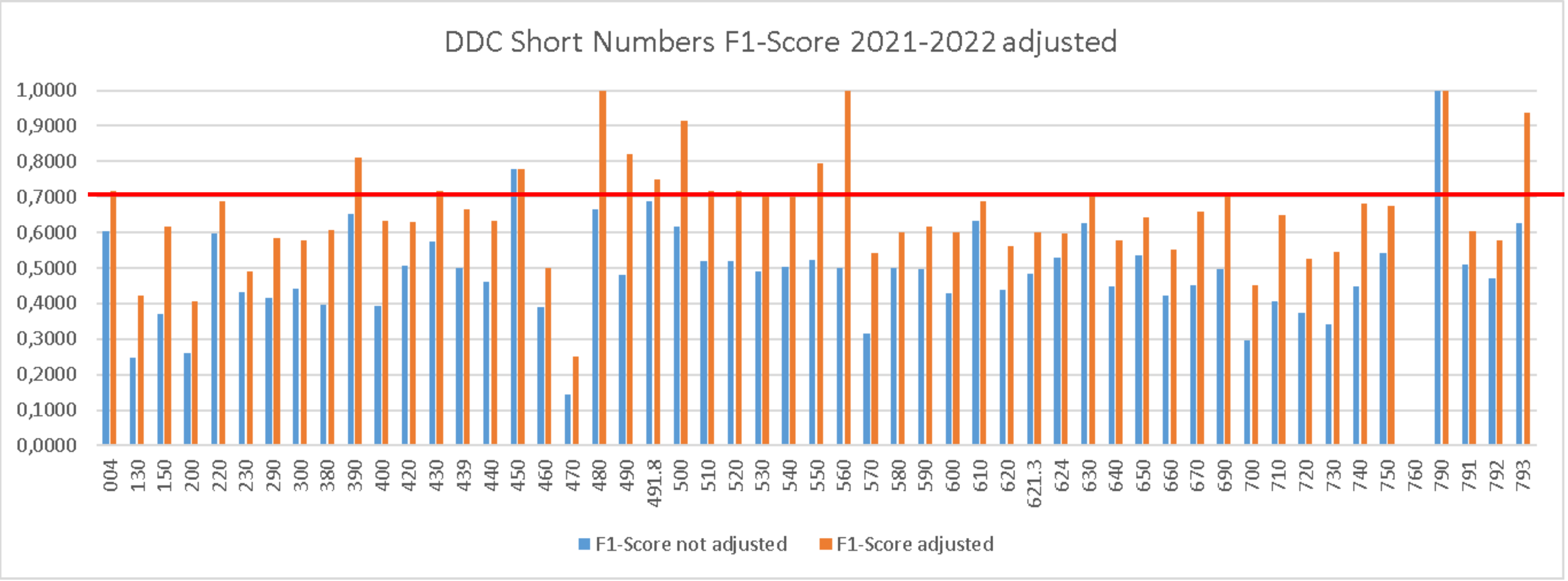
Results DDC Short Numbers 2021-2022



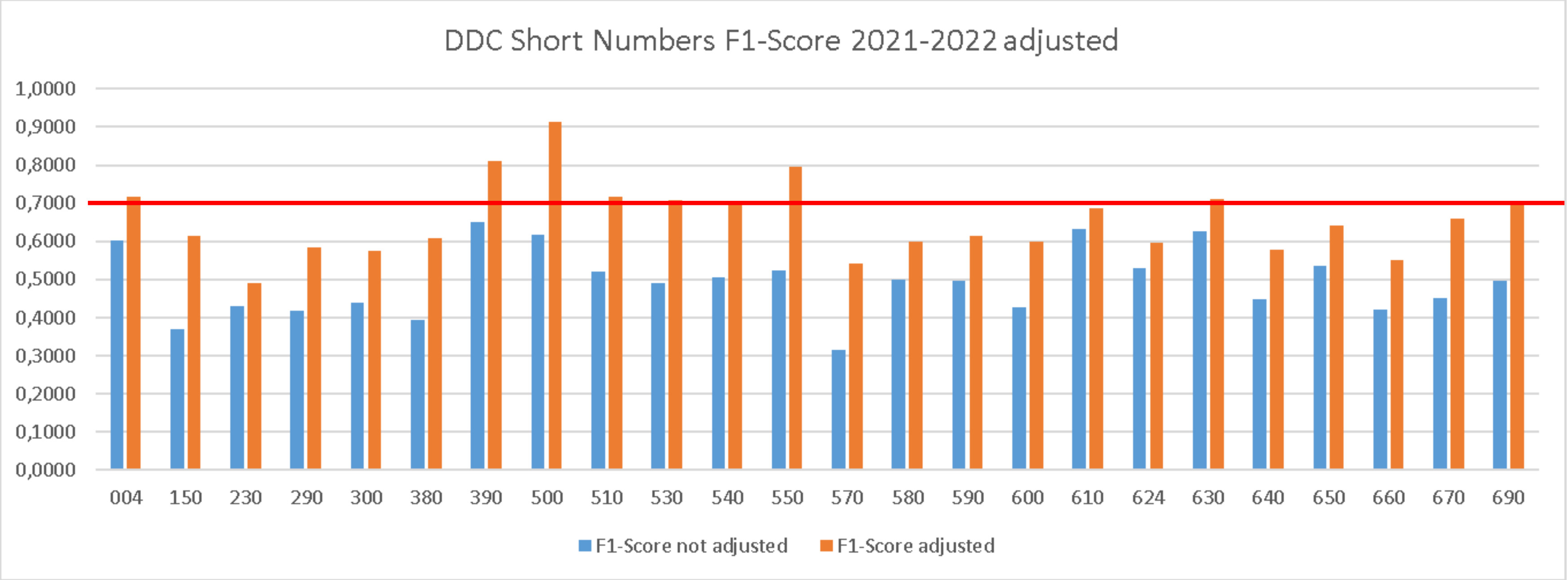
Why are the results worse?

- Less training material
- Higher number of notations
- Error due to predecessor process
  - Result of the machine subject category assignment specifying which machine learning model is to used for assigning DDC short numbers
  - If the first process returns an incorrect result, the subsequent process cannot return a correct result

Results DDC Short Numbers 2021-2022 adjusted



Results DDC Short Numbers 2021-2022 adjusted



Which topics are we working on right now?

- Professionalization of data management
  - Data Version Control (DVC)
  - GitLab
- Revision of the process chain and the machine-learning models used

Thank you for your attention!

Frank Busse
German National Library
Section Automatic Indexing, Online Publications
Adickesallee 1
60322 Frankfurt am Main

mailto:f.busse@dnb.de