DEUTSCHE NATIONAL BIBLIOTHEK

Sandro Uhlmann & Claudia Grote

# Automatic subject indexing with Annif at the German National Library (DNB)

SWIB21
Semantic Web in Libraries

29 Nov - 3 Dec 2021
on the web

Session Controlled Vocabularies
2021-12-01

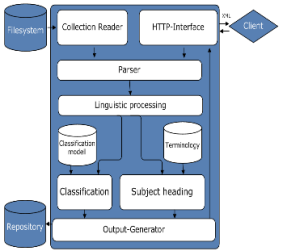
# Automatic cataloguing in the DNB

Automatic classifying of selected online and print publications using DNB subject categories or DDC Short Numbers

(associative approach)

Automatic indexing of selected online and print publications using controlled vocabulary from The Integrated Authority File (GND)

(text mining, lexical approach)




Software:

Averbis Extraction Platform (AEP)

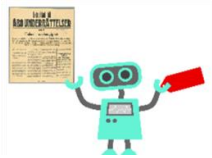
# Initial situation

DEUTSCHE NATIONAL BIBLIOTHEK

- no further development of the automatic cataloguing software DNB-AEP from the provider Averbis
- a new development for the entire system, consisting of DNB web services for control and communication as well as the Averbis software must be planned and implemented
- DNB set up a project called Erschließungsmaschine (EMa) to replace the legacy system with a modular system for automatic subject cataloguing

# Evaluation

annif



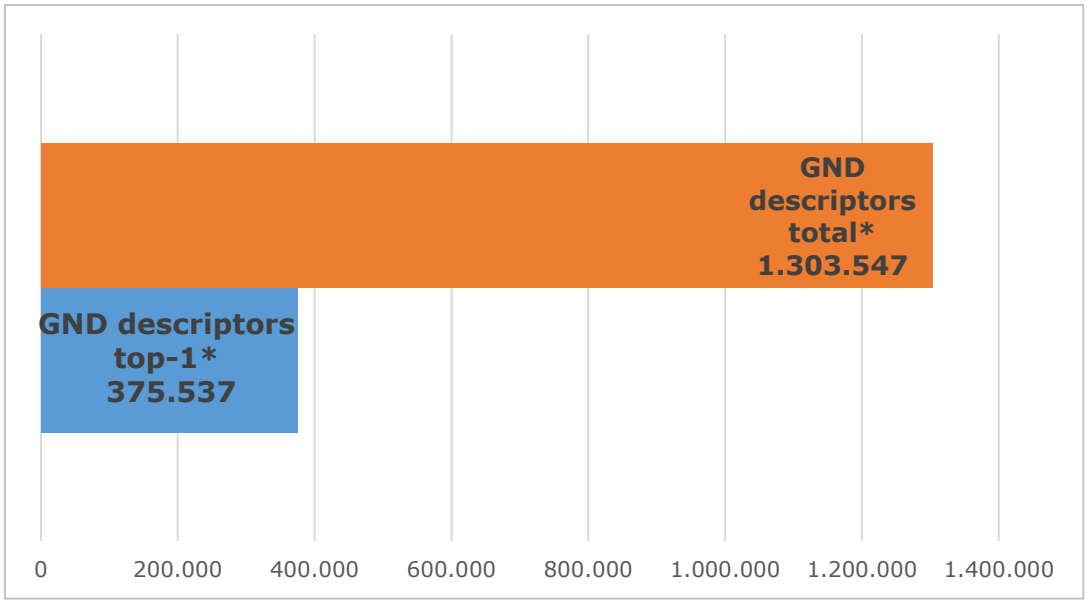
DEUTSCHE NATIONAL BIBLIOTHEK

- developed at the National Library of Finland
- uses different tools for natural language processing & machine learning (associative and lexical approaches) like omikuji, fasttext, MLLM, stwfsa, ...
- is multilingual
- can use any vocabulary in SKOS or simple TSV
- is open source and implemented in Python

Suominen, Osma: Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly, 29 (1) 2019 , pp.1–25. http://doi.org/10.18352
Suominen, Osma; Inkinen, Juho; Lehtinen, Mona: Annif and Finto AI: Developing and Implementing Automated Subject Indexing. PrePrint accepted on 23 June 2021 for publication in JLIS.it https://urn.fi/URN:NBN:fi-fe2021080942632
Github https://github.com/NatLibFi/Annif

# Automatic indexing

GND descriptors approved for subject indexing in the German Integrated Authority File (GND)

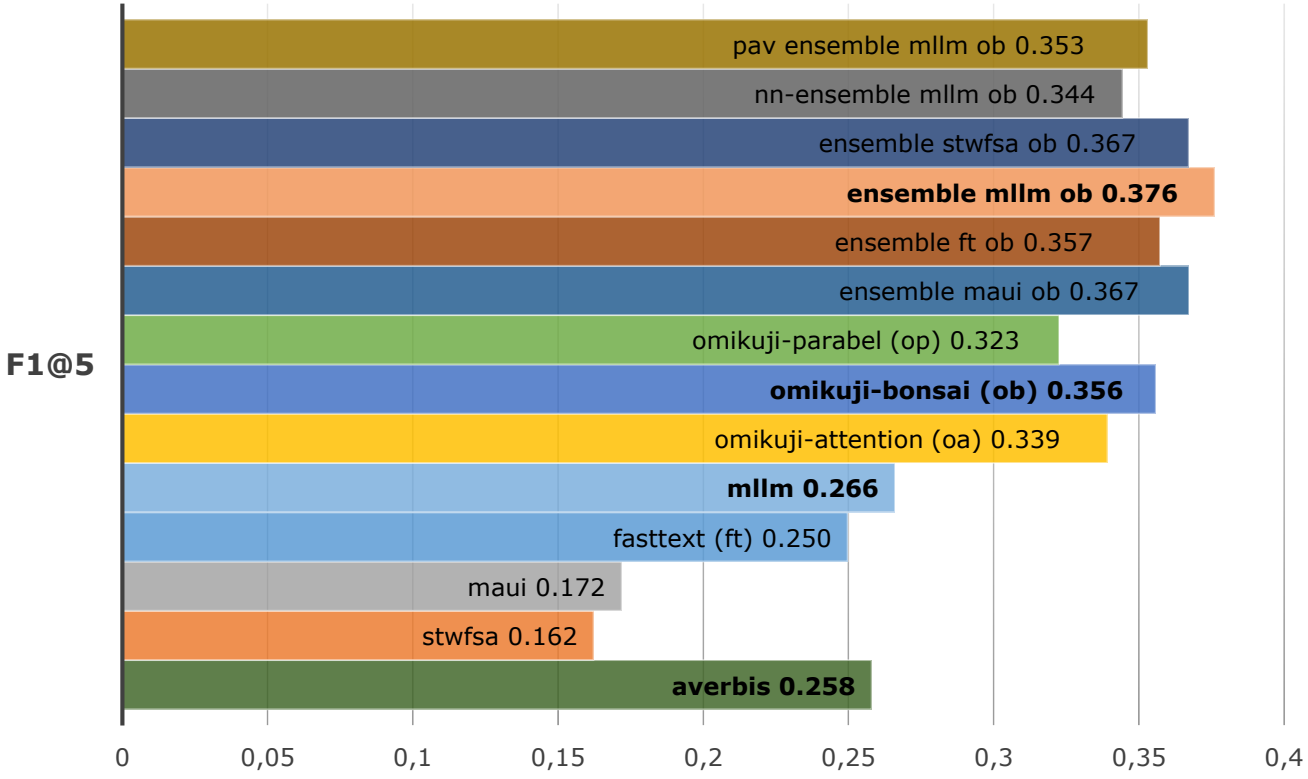



*authority records of subset s assigned to indexing level 1 or indexing level z

Training data, German:

- no complete training set for all GND descriptors
- only 375.537 have at least one text object
- 928,010 have no text object

# Results GND descriptors




Testdata:

1261 Online publications

F1-score

(n=5 descriptors)

# Assessment by subject specialists 

DEUTSCHE NATIONAL BIBLIOTHEK

702 samples (Online publications)

2998 GND descriptors from an ensemble of MLLM & omikuji-bonsai

Assessment of the descriptors on the basis of random samples by human evaluators of the subject cataloging department

Ratings:

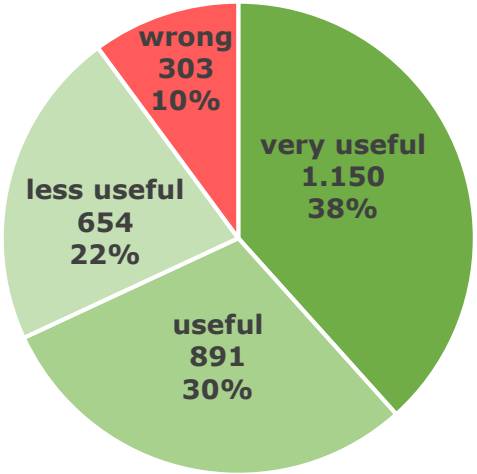
- very useful
- useful
- less useful
- wrong

# Assessment by subject specialists

DEUTSCHE NATIONAL BIBLIOTHEK

702 samples (Online publications)

2998 GND descriptors from an ensemble of MLLM & omikuji-bonsai




GND descriptors n = max 6

1094 Missing aspects

Assessment of the descriptors on the basis of random samples by human evaluators of the subject cataloging department

Ratings:

- very useful
- useful
- less useful
- wrong

# annif for automatic cataloguing @DNB

DEUTSCHE NATIONAL BIBLIOTHEK

3 vocabularies in action

| | GND descriptors | DDC subject categories | DDC short numbers (e.g., medicine) | |
|---|---|---|---|---|
| number of labels | 1,303,547 | 100 | 121 | |
| language | de | de | de | en |
| annif backend | ensemble (omikuji-bonsai + mllm) | omikuji | omikuji | omikuji |
| number of training docs | 2,415,549 + 37,001 | 473,000 | 66,700 | 10,900 |
| model size on disk | 13 GB + 984 MB | 21 GB | 3.3 GB | 716 MB |

# annif @DNB: modelling vs. cataloguing

DEUTSCHE NATIONAL BIBLIOTHEK

Evaluation environment

- task: model training/validation/test, corpus management
- Annif usage: CLI
- CPU: 16 cores @2.60 GHz
- RAM: 640 GB
- OS: Ubuntu Linux

# annif @DNB: modelling vs. cataloguing

DEUTSCHE NATIONAL BIBLIOTHEK

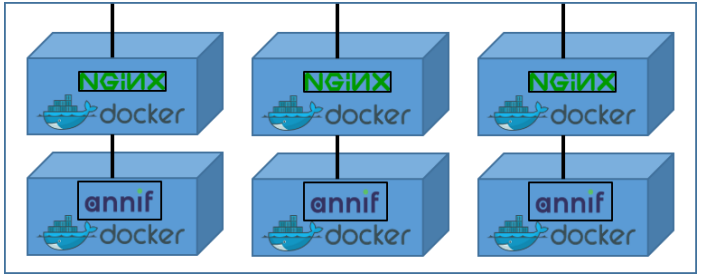
Evaluation environment

- task: model training/validation/test, corpus management
- Annif usage: CLI
- CPU: 16 cores @2.60 GHz
- RAM: 640 GB
- OS: Ubuntu Linux

```
$ annif [OPTIONS] COMMAND [ARGS] ...
```

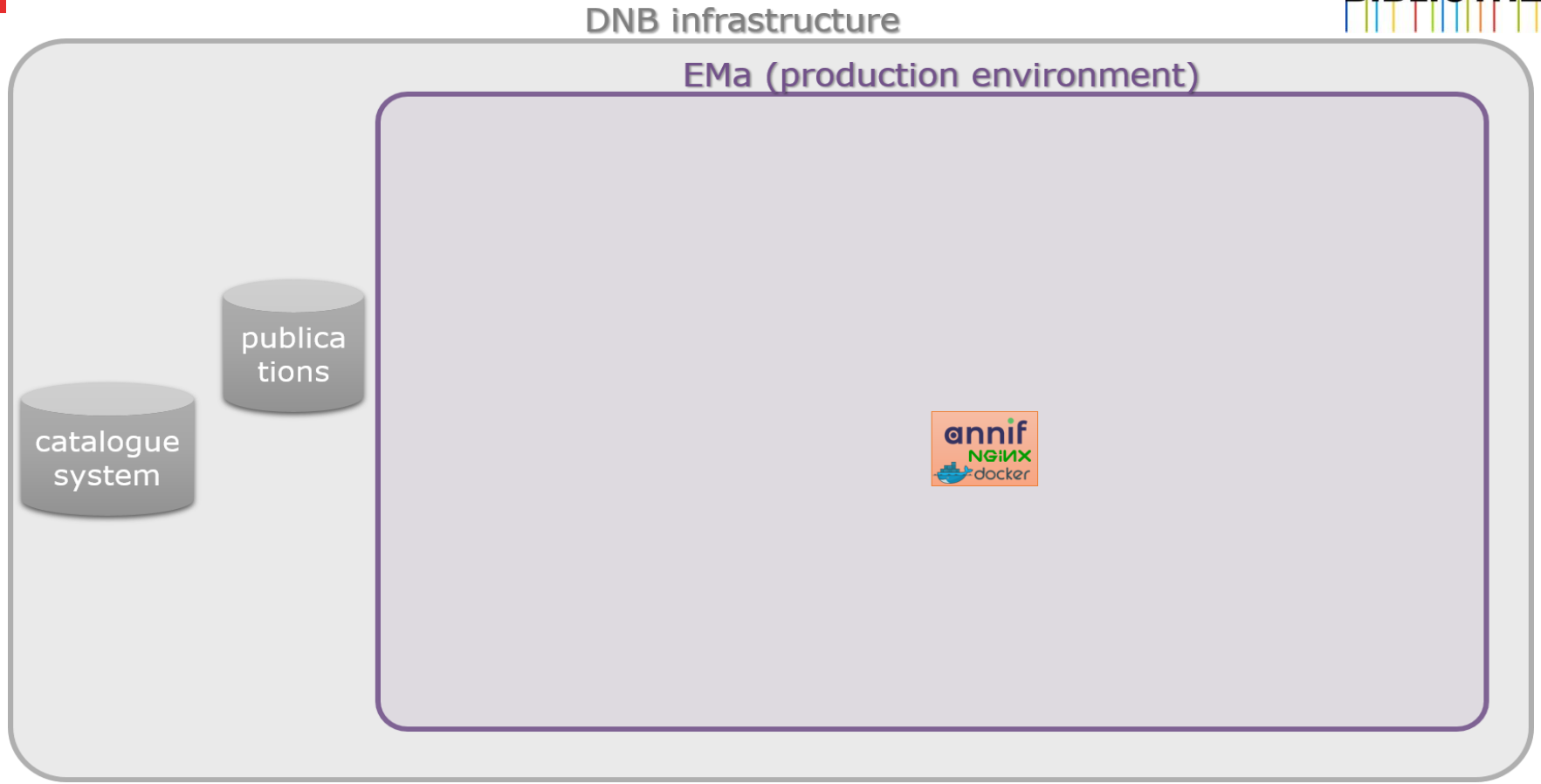
EMa production environment

- task: metadata generation for automatic cataloguing
- Annif usage: 3 x REST service with Annif and NGINX, each run in a Docker container, configured with docker-compose
- CPU: 8 cores @2.60 GHz
- RAM: 128 GB
- OS: Ubuntu Linux



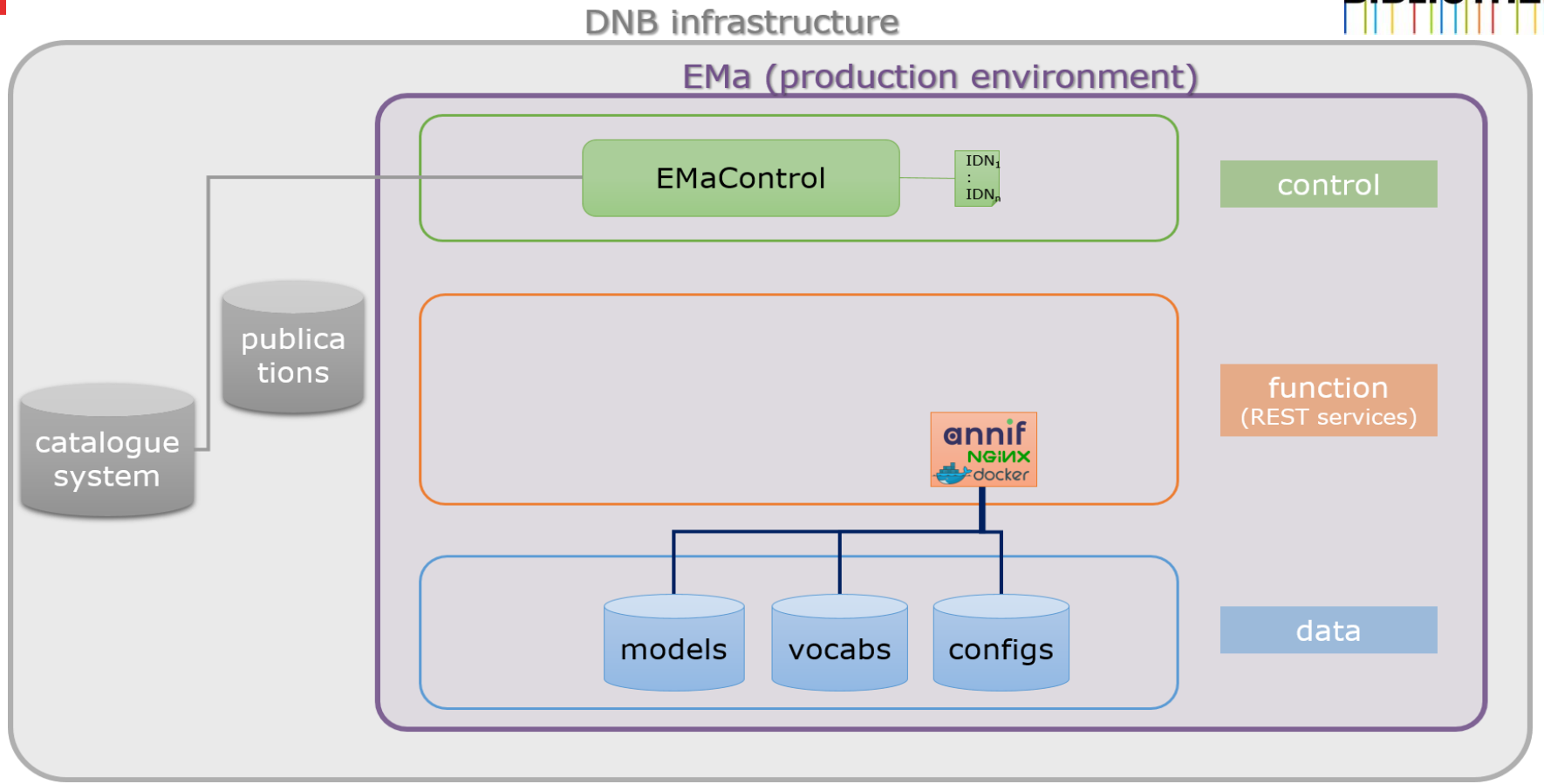
# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



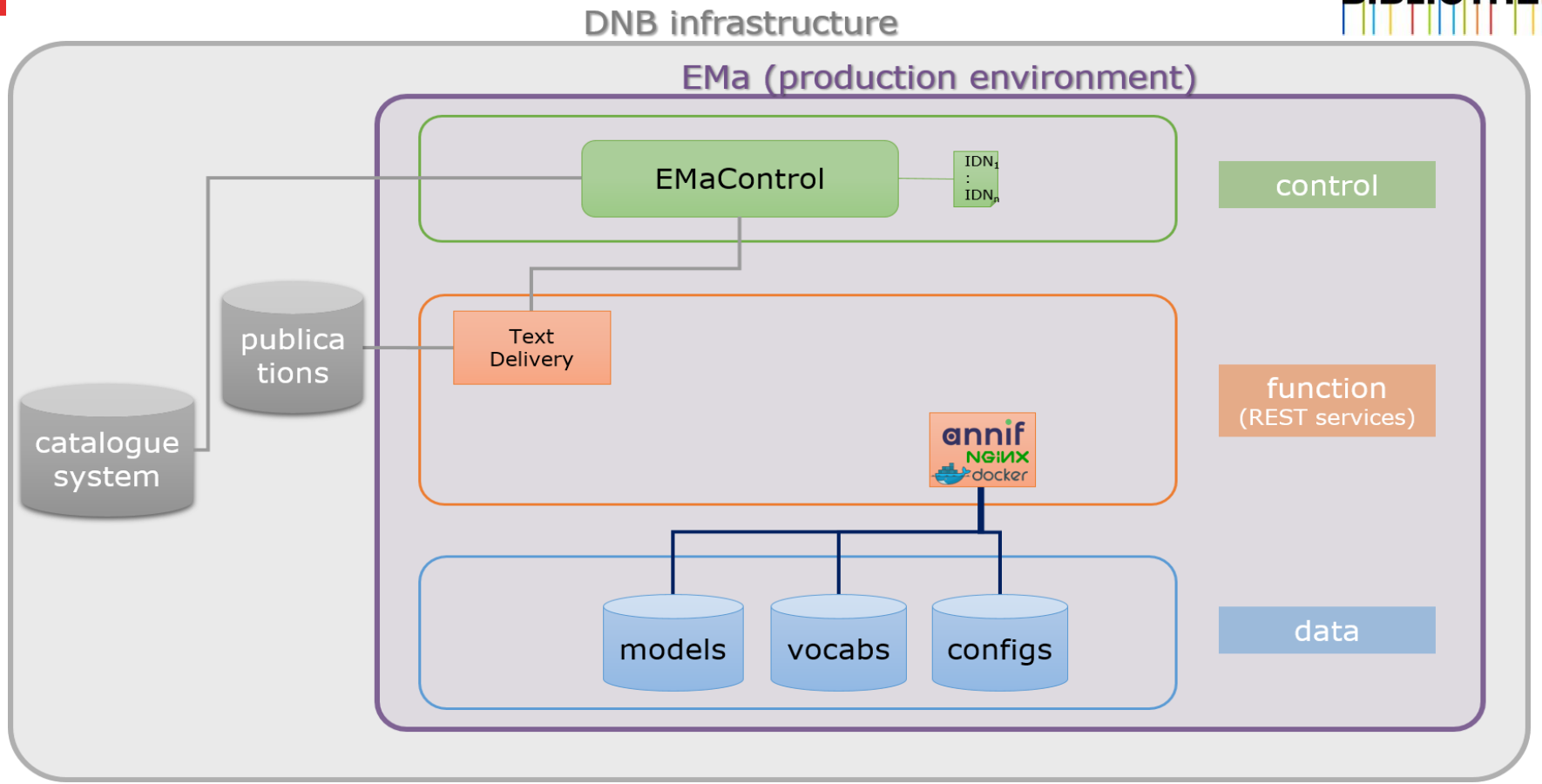
# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



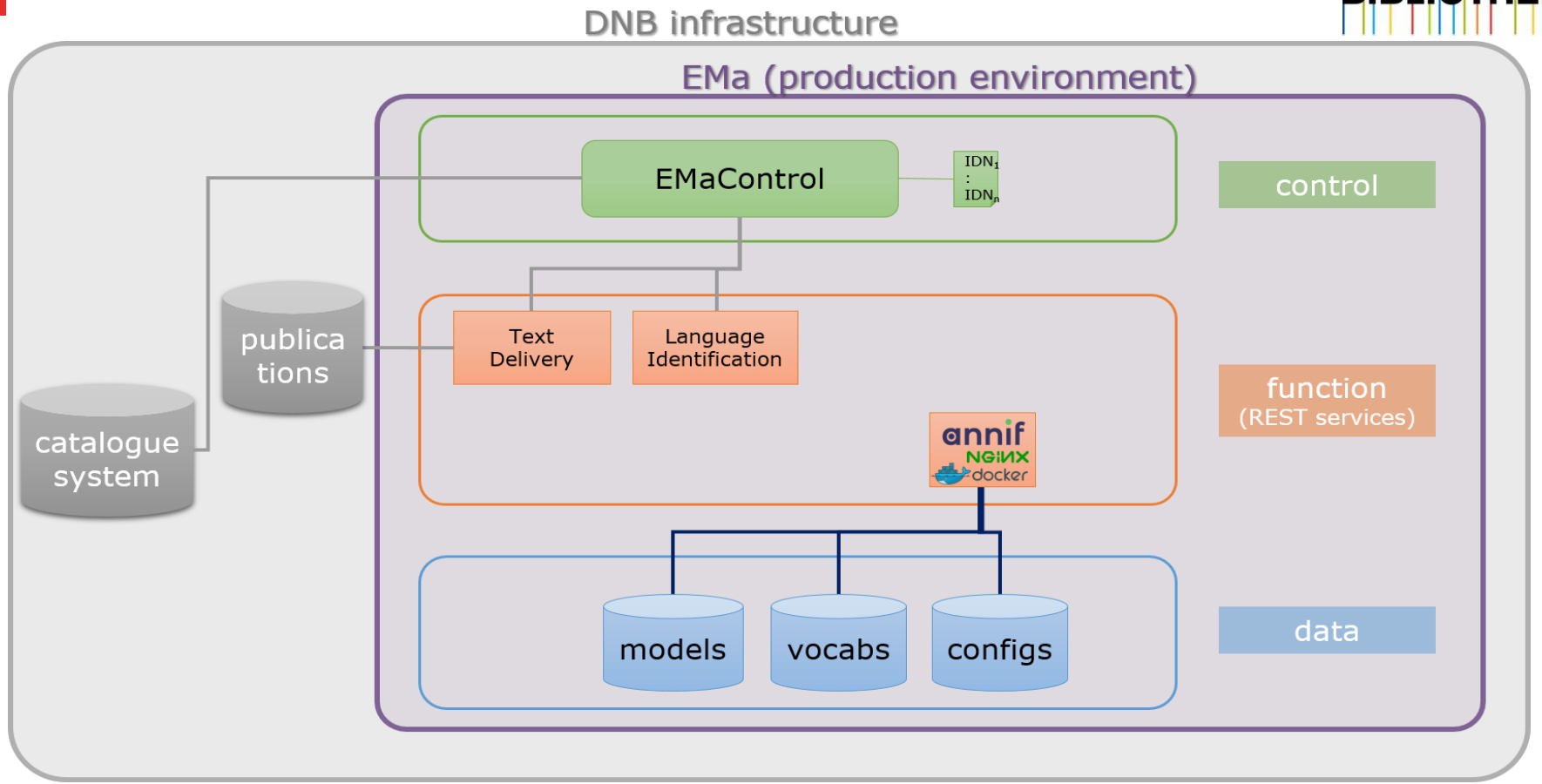
# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



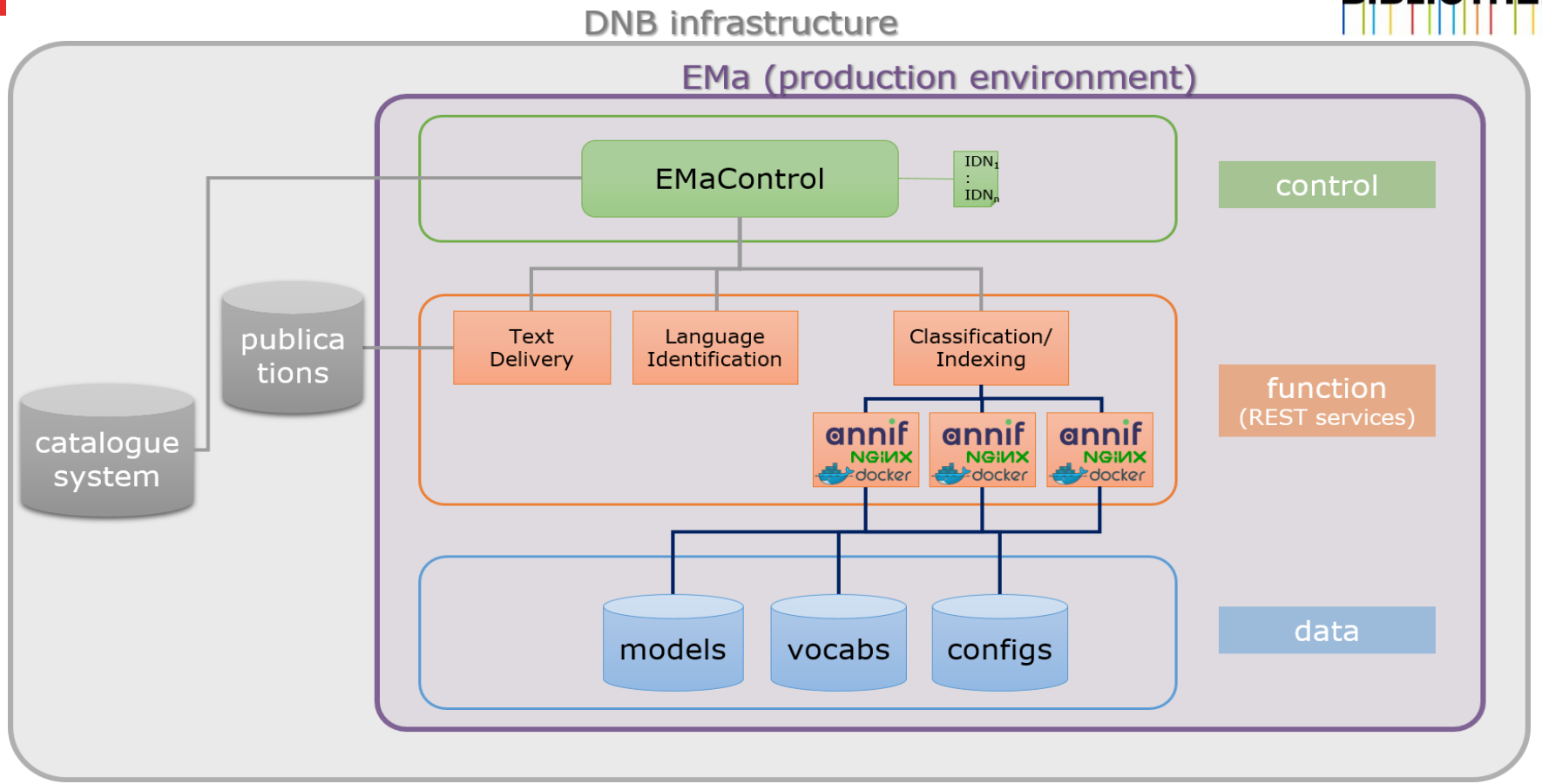
# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



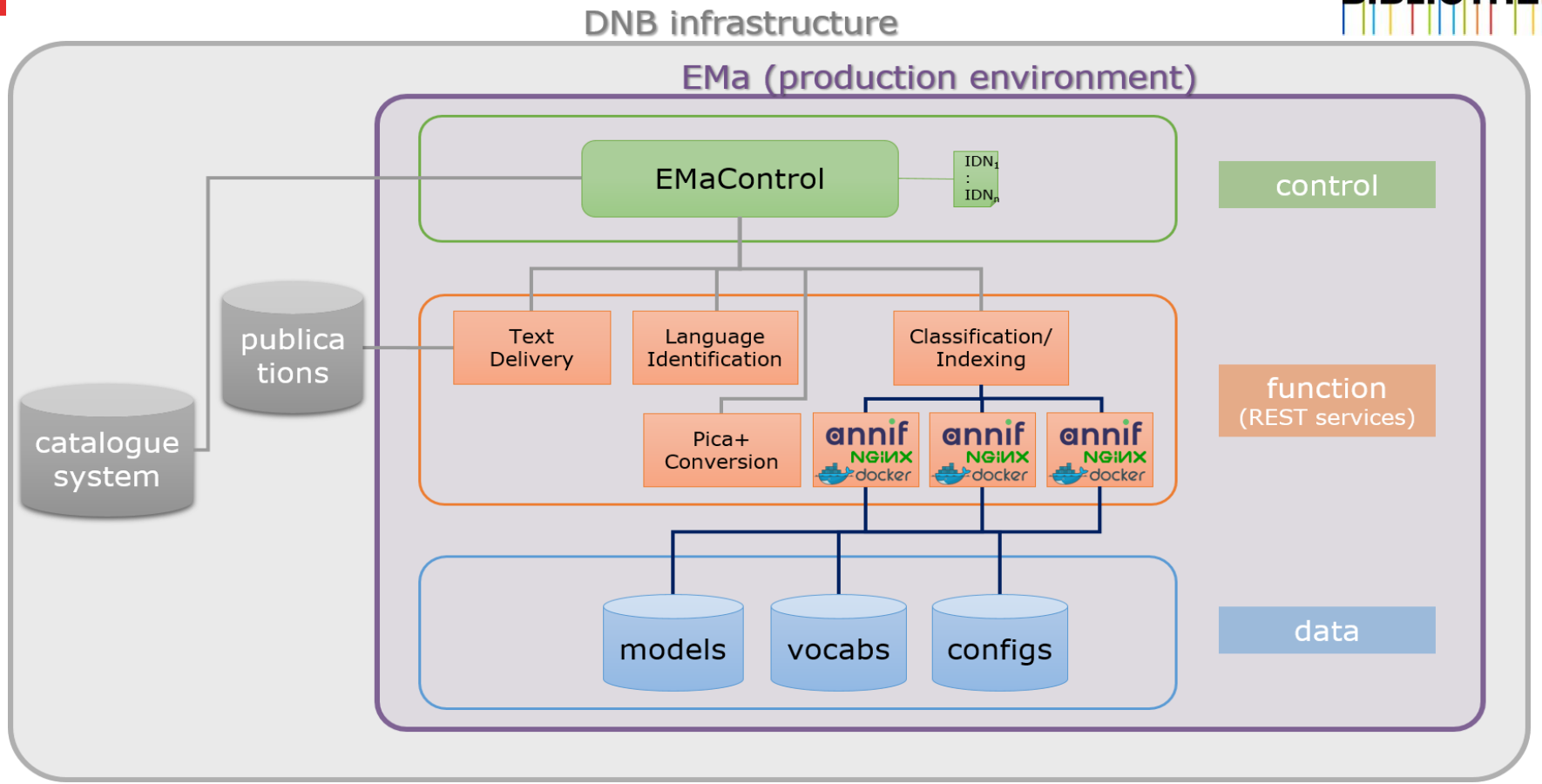
# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



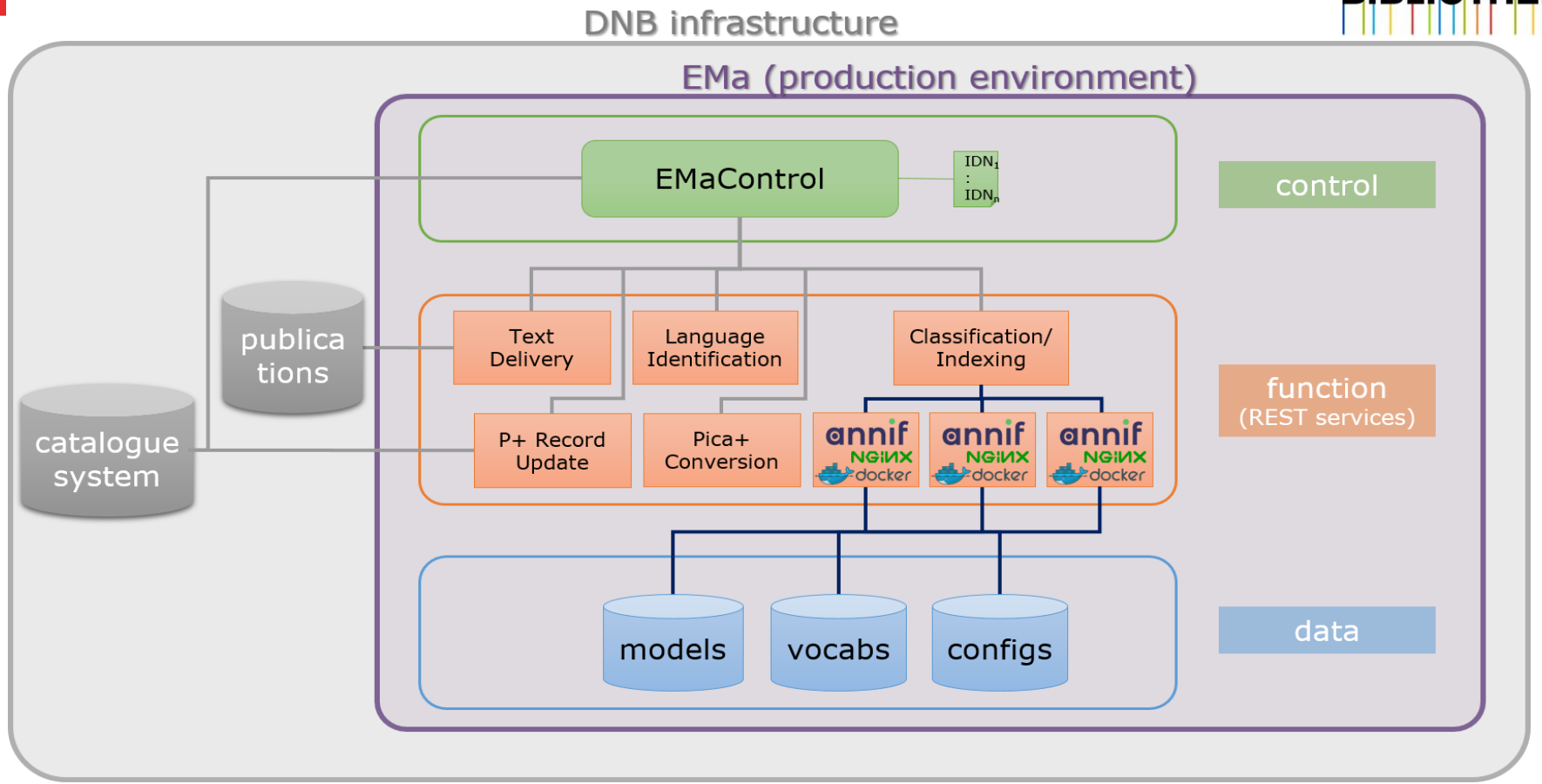
# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



# annif system integration @DNB (architecture sketch)

DEUTSCHE NATIONAL BIBLIOTHEK



# Outlook

Until March 2022:

- Test the process flows of all services
- Go-live at least with automatic indexing (GND descriptors) for German-language online publications

From April 2022:

- Transfer of all topics into routine and further development in an agile and iterative way of working
- Retirement of the legacy system will be the focus for the transition from project status to product status
- Text optimization: improve content fidelity/reliability of textual representations of digital publications as a basis for Automatic subject cataloguing

DEUTSCHE NATIONAL BIBLIOTHEK

# Thank you for your attention.

s.uhlmann@dnb.de | c.grote@dnb.de

*There are a few more slides with additional information about Annif results with DDC subject categories, DDC short numbers as well as provenance data in the German National Library, which were not presented at SWIB21 today, but might be of interest.*

**Recommendation** 💡 **pica-rs** - tool to work with bibliographic records encoded in PICA+ in a fast and efficient way
https://github.com/deutsche-nationalbibliothek/pica-rs

# Automatic classifying

DEUTSCHE NATIONAL BIBLIOTHEK

## DDC Subject Categories & DDC Short Numbers

**DDC Subject Categories and included DDC classes for the New Release Service and the Series A, B, C, H and O of the Deutsche Nationalbibliografie (based on DDC 23)**

| Class | Discipline | Included DDC Classes |
|---|---|---|
| ***000*** | ***Generalities, computers, information*** | |
| 000 | Generalities | 000-003 |
| 004 | Computer science | 004-006 |
| 010 | Bibliography | 010 |
| 020 | Library and information sciences | 020 |
| 030 | Encyclopedic works | 030 |
| 050 | Magazines, journals and serials | 050 |
| 060 | Organizations and museology | 060 |
| 070 | News media, journalism, publishing | 070 |
| 080 | General collections | 080 |
| 090 | Manuscripts and rare books | 090 |
| ***100*** | ***Philosophy and psychology*** | |
| 100 | Philosophy | 100-120, 140, 160-190 |
| 130 | Parapsychology, occultism | 130 |
| 150 | Psychology | 150 |
| ***200*** | ***Religion*** | |
| 200 | Religion, philosophy and theory of religion | 200, 210 |
| 220 | Bible | 220 |
| 230 | Theology, Christianity | 230-280 |
| 290 | Other religions | 290 |
| ***300*** | ***Social sciences*** | |
| 300 | Social sciences, sociology, anthropology | 300 |
| 310 | General statistics | 310 |
| 320 | Political science | 320 |

| | |
|---|---|
| DDC Subject Category | 610 |
| Full DDC Number | 618.92398 |
| DDC Short Number | 618.92 |

**618.92398 Pediatrics Adipositas**

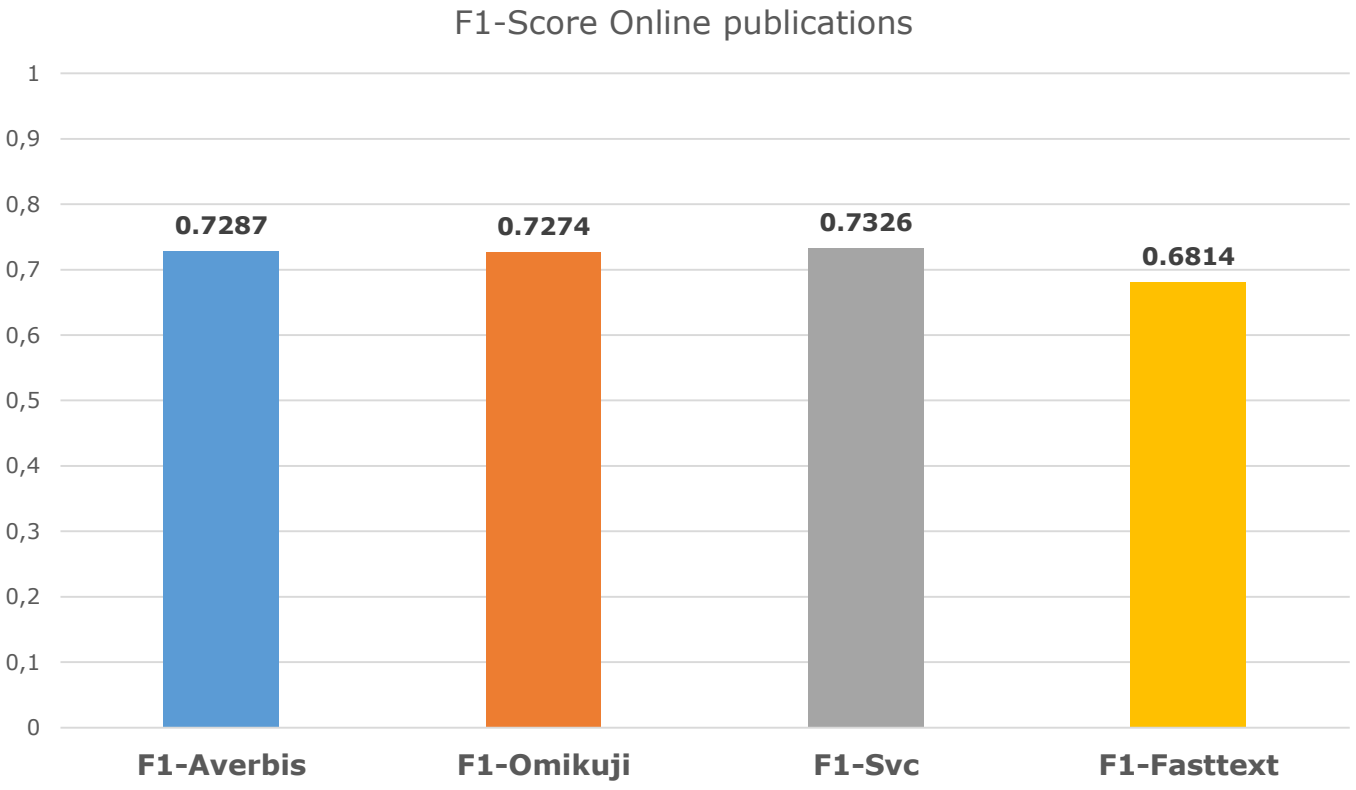
| | |
|---|---|
| DDC Subject Category | 300 |
| Full DDC Number | 303.625088297 0956 |
| DDC Short Number | 303.6 |

**303.625 Terrorism Muslims Middle East**
**303.6 Conflict and conflict resolution**

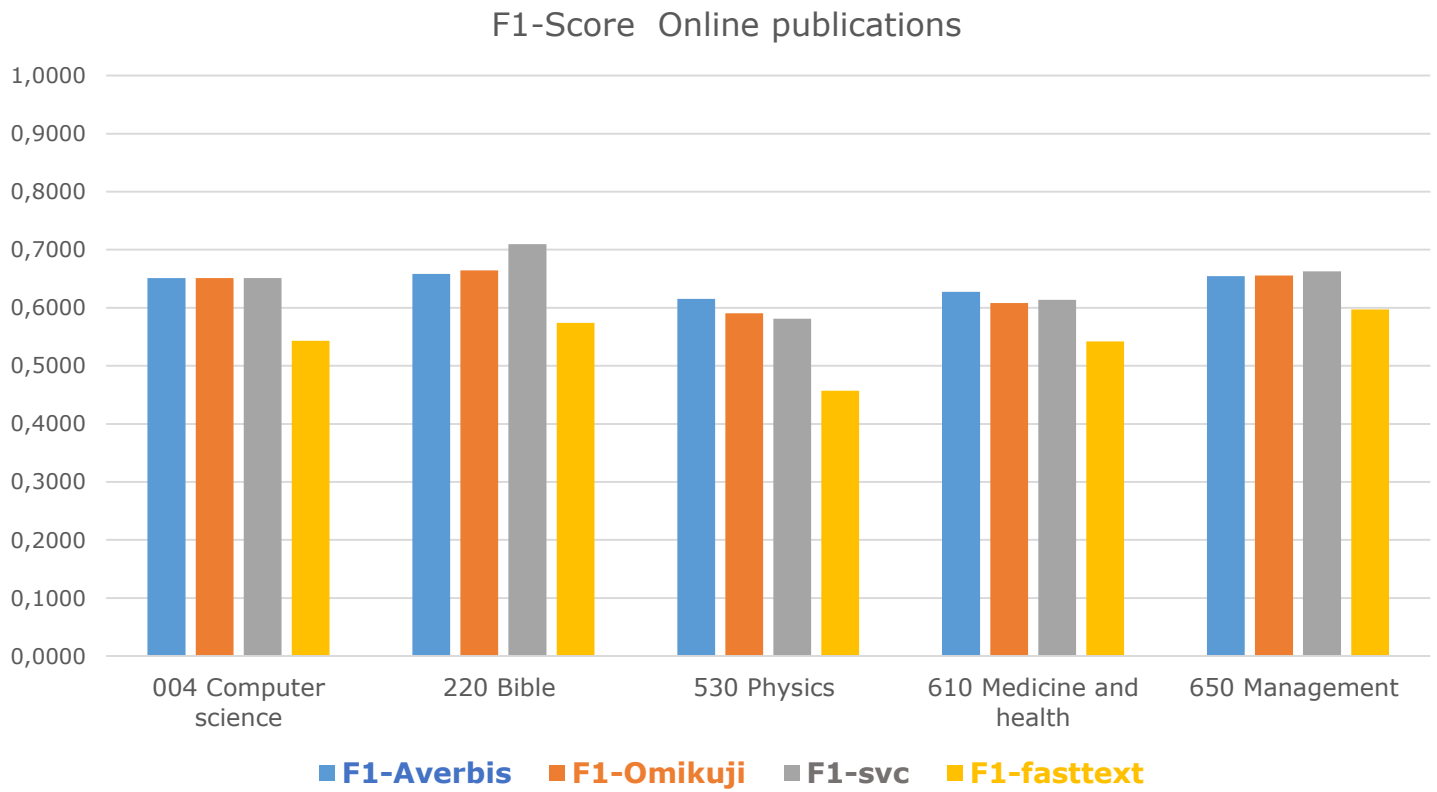
| | |
|---|---|
| DDC Subject Category | 640 |
| Full DDC Number | 641.563620954165 |
| DDC Short Number | 641.56362 |

**641.563620954165**
**Vegan dishes from Nagaland (India)**

# Results - DDC Subject Categories

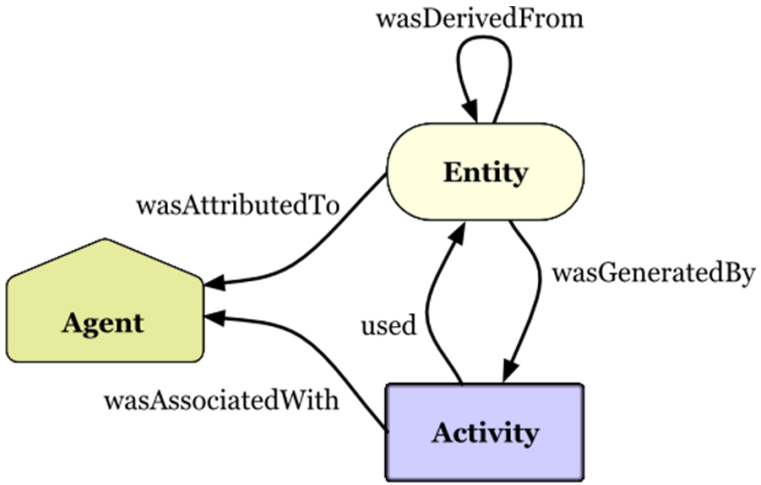


# Results - DDC Short Numbers



# Metadata provenance

*Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness.*




https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

# Metadata provenance in Pica+, Marc21 & LOD

DEUTSCHE NATIONAL BIBLIOTHEK

```
021A $a Feelgood Management $d Mit Wertschätzung und Menschlichkeit erfolgreich in die Arbeitswelt von morgen
```

## Pica+

*machine generated metadata*

*gnd descriptor generated by an activity from agent Averbis Extraction Platform*

```
044H $b GND $9 040372782 $7 Tsz $V saz $A gnd $0 4037278-9 $a Management $E m $H aepgnd $K 0,15344 $D 2021-06-30
044H $b GND $9 040261476 $7 Ts1 $V saz $A gnd $0 4026147-5 $a Humanität $E m $H aepgnd $K 0,07171 $D 2021-06-30
```

## Marc21

```
650  7$83\p$0(DE-588)4037278-9$0https://d-nb.info/gnd/4037278-9$0(DE-101)040372782$aManagement$2gnd
883 0 $83\p$aaepgnd$c0,15344$d20210630$qDE-101$uhttps://d-nb.info/provenance/plan#aepgnd
```

*gnd descriptor generated by an activity from agent Averbis Extraction Platform*

*provenance plan*

## Turtle

```
_:node1fj6210jqx215 agrelon:metadataConfidence "0.15344"^^xsd:float;
  prov:generatedAtTime "2021-06-30"^^xsd:date;
  a prov:Entity, dnbt:QualifiedSubject;
  dnbt:qualifiesSubject <https://d-nb.info/gnd/4037278-9>;
  prov:wasGeneratedBy dnb-prov-a:m:2021-06-30 .
```