Insight into the machine-based subject cataloguing at the German National Library

Christoph Poley, DNB AEN

@swib2022
Outline

- Introduction
- Workflows
  - Productive Workflow of EMa
  - Data management workflow of text corpora
  - Workflow of GND vocabulary
  - Evaluation workflow
- Resources
- The DNB AI project
Introduction
German National Library

- Central archival library
- Was founded in 1913 in Leipzig – Two locations now
- Law Regarding the German National Library (DNBG)
- Collect everything that is published in Germany: Two hardcopies and Online Publication
Situation: Increasing number of online publications

7500 online publications each day

How to handle that?

1st EMa release: spring 2022
Automatic indexing: Use cases
Heart of DNB Erschließungsmaschine: **annif**

- Open source toolbox 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, ...
- Works multilingual
- Uses standard interfaces and formats
- Growing international community
Workflows
Productive Workflow of EMa

daily, fully automatically

text delivery service
- catalogue management system
- PDF/EPUB extraction

text language detection service
- Apache Tika

classification & indexing service
- Annif
- System y

Pica cataloguing service
- Catalogue management system
Data management workflow of text corpora

DVC

- gold standard: Pica, pica_rs
- split: Python/R
- fetch texts: Python/curl
- cleanup: Python

Python libraries: annif, csv, csvs, tsv, text, curl
Workflow of GND vocabulary
Evaluation workflow

- annotate
  - pica
  - WinIBW
- extract
  - pica
  - pica_rs
- analyse
  - csv
  - Python, ...

- very useful
- useful
- less useful
- wrong
Intellectual and automatic indexing and subject cataloguing: A cycle

WinIBW

intellectual evaluation

subject cataloguing „gold standard“

productive workflow of EMa

data management workflow

„human in the loop“
Resources
Hardware for machine-based subject cataloguing

- Usage of VMWare infrastructure
- Annif in production:
  - 3 identical Stages: test, approval, productive
  - CPU: 8 cores, Intel® Xeon® Platinum 8260 @ 2.4 GHz
  - RAM: 128 G
  - HDD: 250G via NetApp
- Data management workflow (train vocabulary):
  - CPU: 16 cores, Intel® Xeon® E5-2690 v4 @ 2.6 GHz
  - RAM: 640 G
  - HDD: 4T via NetApp
Recommanded profession skills

- Librarian, data scientist: develop / maintain AI models, quality assurance
- System architect, Software developer: build a productive system / fill gaps in the workflow
- System administrators, hardware architects: build and maintain hardware environment
- Data Analyst: Vocabularies, corpora, metrics ...
- ...
Conclusions to the usage of machine-based suggestions

Usage of machine based solutions ... 
- is often more than a side job
- combines different profession skills
- needs special hardware
- needs and ties long-term manpower.

➢ We are dealing with Services and not projects only.

They have to be an part of libraries strategies and staffing.
The DNB AI project
The DNB AI project

Subject cataloguing at the German National Library using AI methods

- Project duration: 4 years
- Work start: October 2021 – March 2025
- Funded as part of the AI Strategy of the Federal Government of Germany
Purpose of the DNB AI Project 1/2

- **Improving quality** of automated subject cataloguing
- Exploring/testing a wide range of **innovative methods**
- **Proper representation** of **GND** data – Preparing the vocabulary
- Making **better use** of the **potential** of the **GND**
Purpose of the DNB AI Project 2/2

- **Concepts without GND representation** should also be recognized
- Provide suitable **new tools** for practical use
- Expanding **AI competencies in cultural institutions**
Thank you for your attention!

Christoph Poley
German National Library
Phone: +49 341 2271-247
mail: c.poley@dnb.de
http://www.dnb.de