Christoph Poley, DNB AEN

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

@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

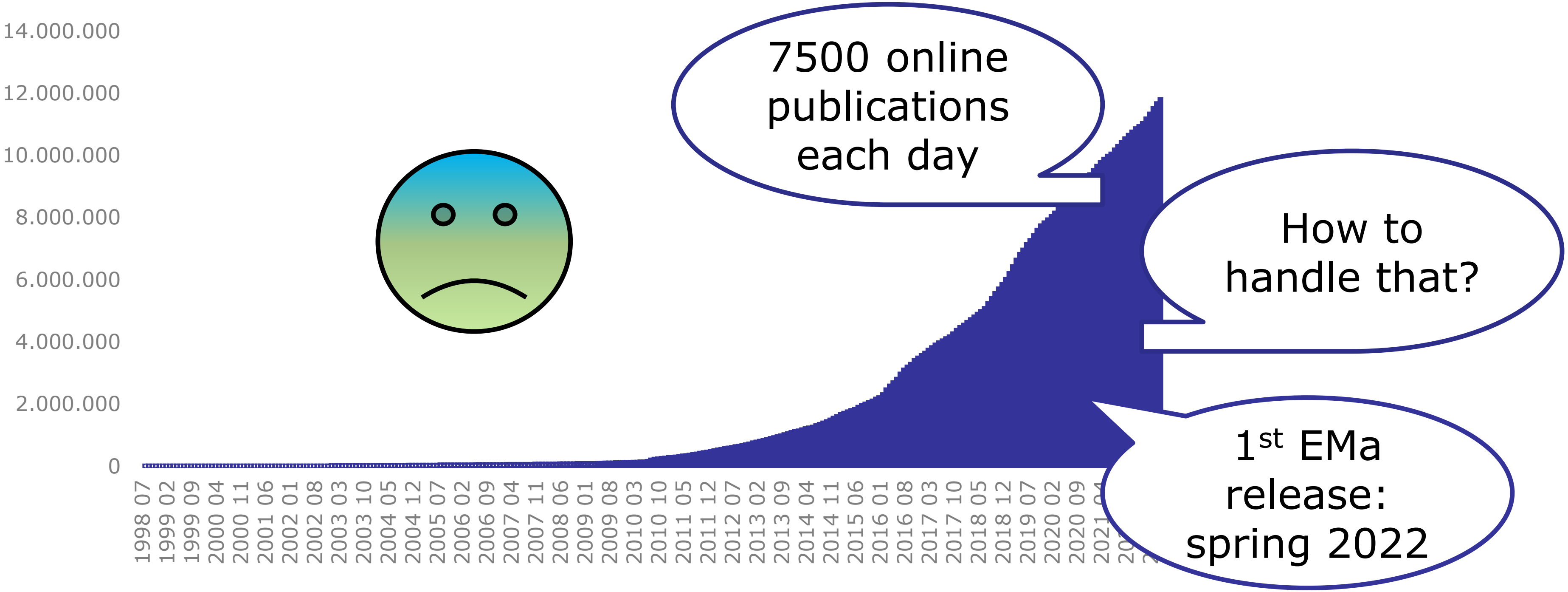
# German National Library

Frankfurt

Leipzig

- 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

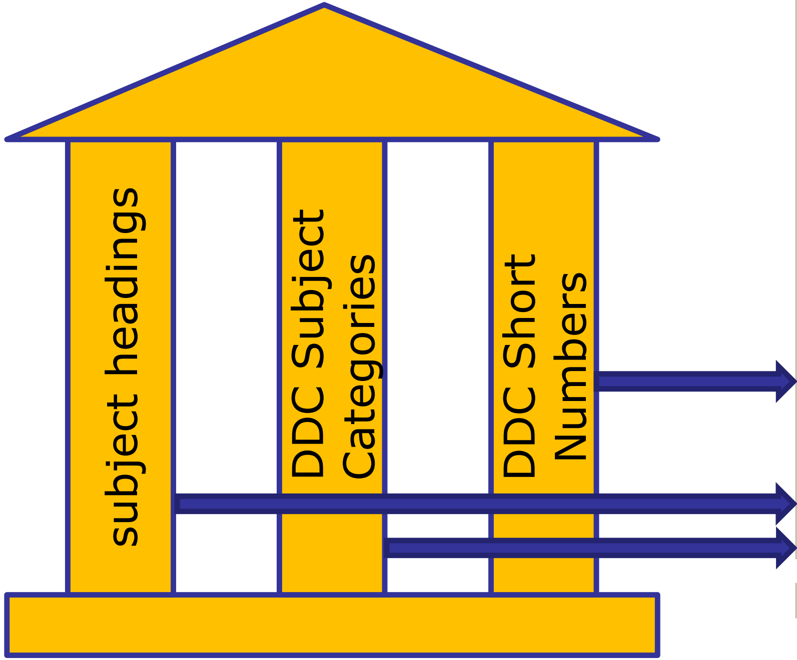


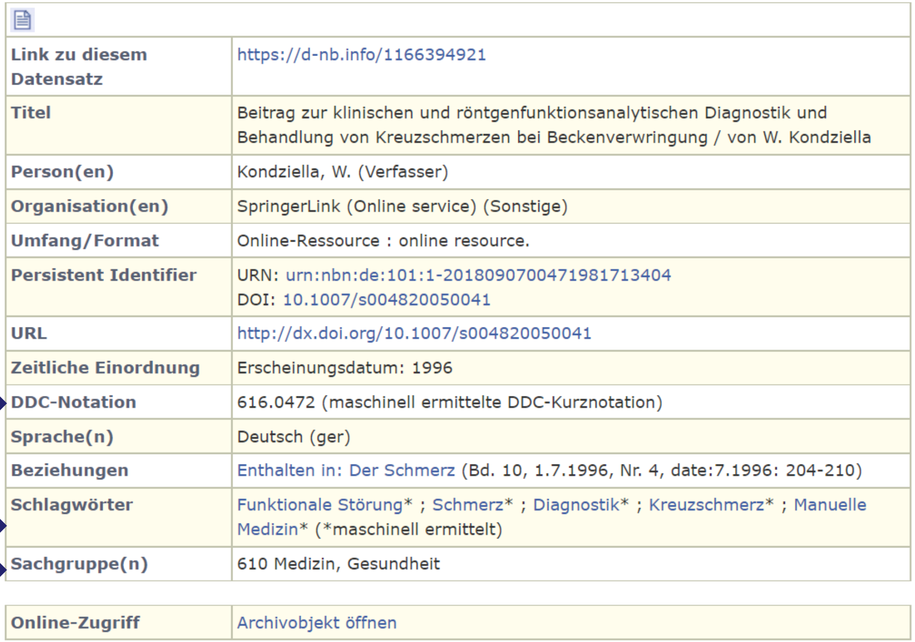
7500 online publications each day

How to handle that?

1st EMa release: spring 2022

# Automatic indexing: Use cases

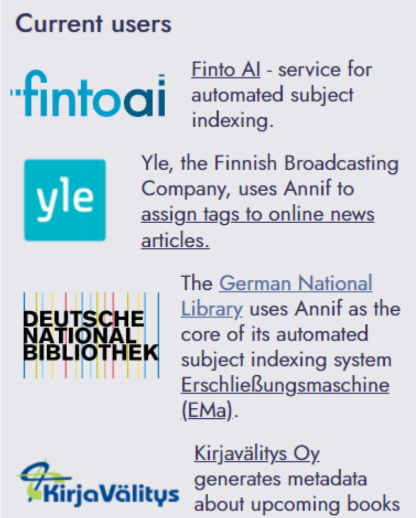




| | |
|---|---|
| **Link zu diesem Datensatz** | https://d-nb.info/1166394921 |
| **Titel** | Beitrag zur klinischen und röntgenfunktionsanalytischen Diagnostik und Behandlung von Kreuzschmerzen bei Beckenverwringung / von W. Kondziella |
| **Person(en)** | Kondziella, W. (Verfasser) |
| **Organisation(en)** | SpringerLink (Online service) (Sonstige) |
| **Umfang/Format** | Online-Ressource : online resource. |
| **Persistent Identifier** | URN: urn:nbn:de:101:1-2018090700471981713404<br>DOI: 10.1007/s004820050041 |
| **URL** | http://dx.doi.org/10.1007/s004820050041 |
| **Zeitliche Einordnung** | Erscheinungsdatum: 1996 |
| **DDC-Notation** | 616.0472 (maschinell ermittelte DDC-Kurznotation) |
| **Sprache(n)** | Deutsch (ger) |
| **Beziehungen** | Enthalten in: Der Schmerz (Bd. 10, 1.7.1996, Nr. 4, date:7.1996: 204-210) |
| **Schlagwörter** | Funktionale Störung* ; Schmerz* ; Diagnostik* ; Kreuzschmerz* ; Manuelle Medizin* (*maschinell ermittelt) |
| **Sachgruppe(n)** | 610 Medizin, Gesundheit |
| **Online-Zugriff** | Archivobjekt öffnen |

# 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



Current users

Finto AI - service for automated subject indexing.

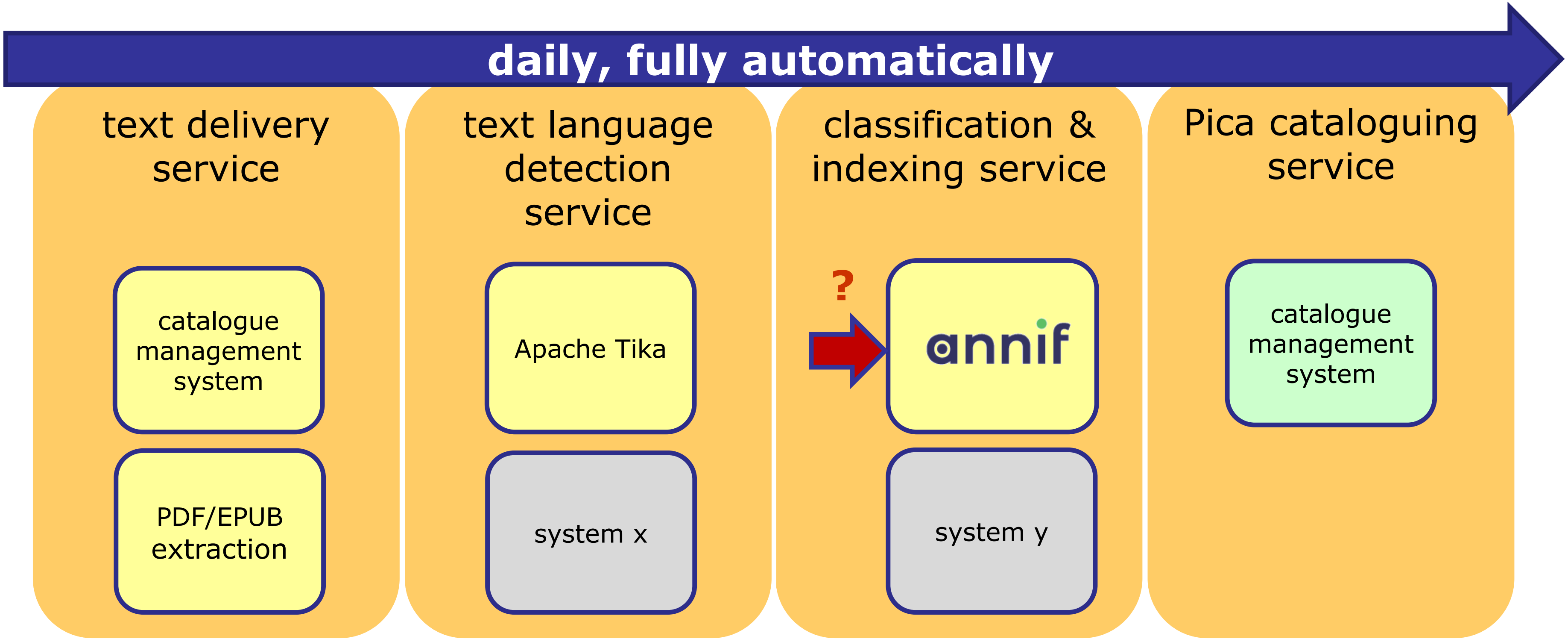
Yle, the Finnish Broadcasting Company, uses Annif to assign tags to online news articles.

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

Kirjavälitys Oy generates metadata about upcoming books

# Workflows

# Productive Workflow of EMa



daily, fully automatically

text delivery service

catalogue management system

PDF/EPUB extraction

text language detection service

Apache Tika

system x

classification & indexing service

?

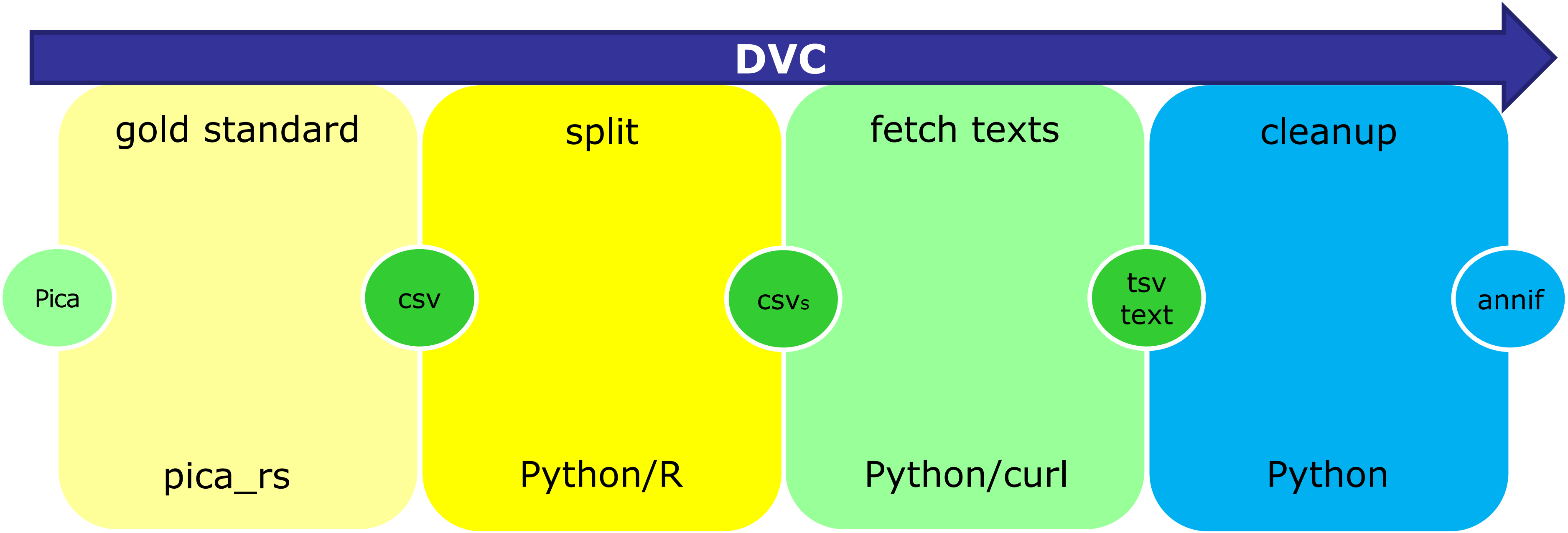
annif

system y

Pica cataloguing service

catalogue management system

# Data management workflow of text corpora



DVC

Pica

gold standard

pica_rs

csv

split

Python/R

csvs

fetch texts

Python/curl

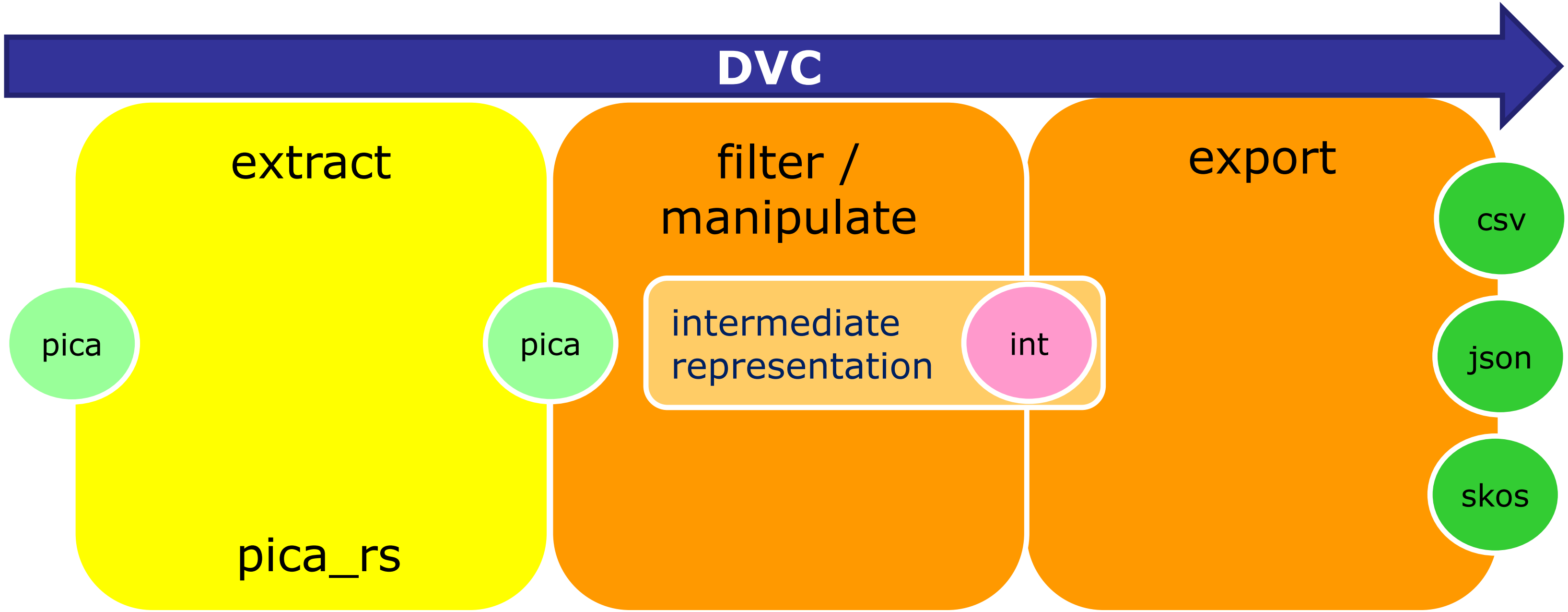
tsv
text

cleanup

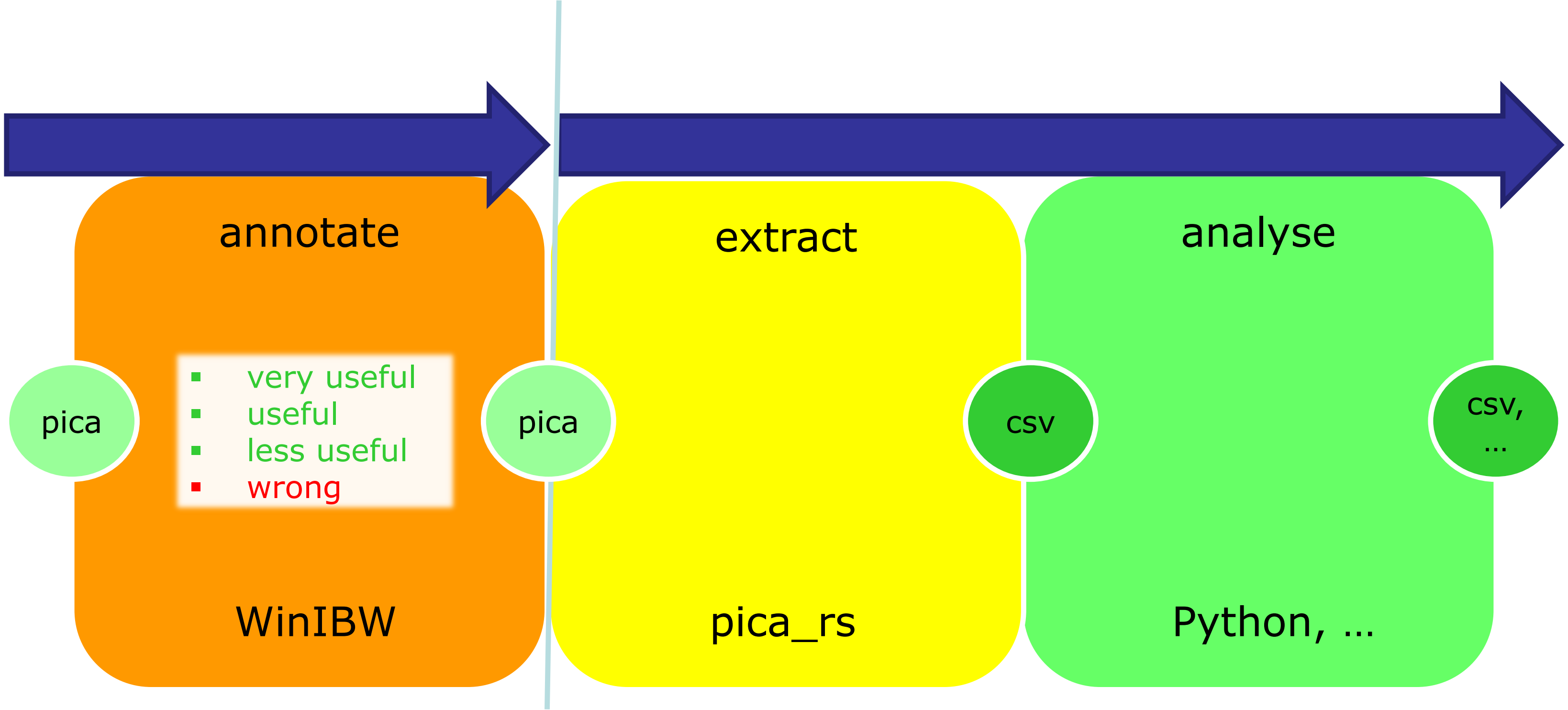
Python

annif

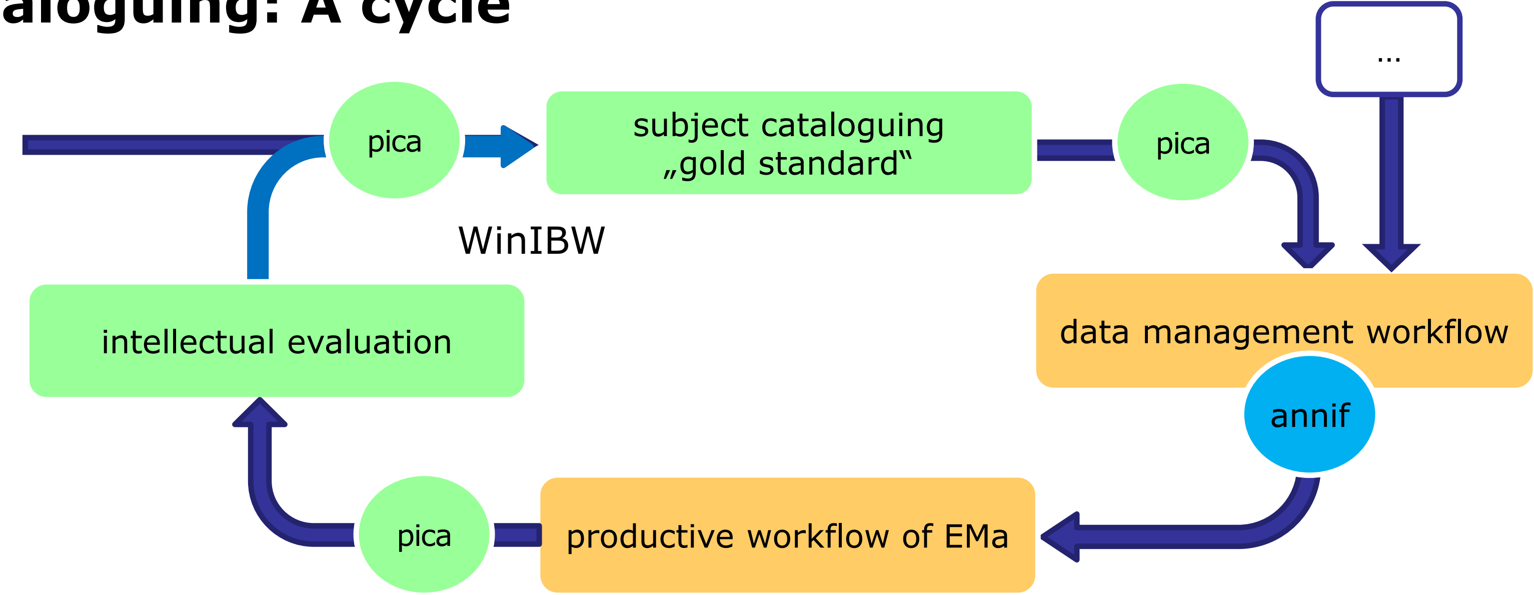
# Workflow of GND vocabulary



# Evaluation workflow



# Intellectual and automatic indexing and subject cataloguing: A cycle



„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 vocalulary):
  - 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