

Nico Wagner, Jan-Helge Jacobs | Deutsche Nationalbibliothek, Leipzig

Aufbereitung der GND für KI-Verfahren

Inhaltsverzeichnis

- 1. Einführung**
- 2. Herausforderungen**
- 3. Anforderungen**
- 4. Demo „gnd-rs“**
- 5. Ausblick**

Einführung

- Bereitstellung der GND im SKOS-Format für Annif
 - Einschränkung auf Entitäten, die den Minimalanforderungen zur Sacherschließung genügen
 - Erzeugung von Literalen und Relationen aus der Stammdatei (PICA+)
 - Verwendung von internen Identifikatoren als *Internationalized Resource Identifier* (IRI)

- Bereitstellung der GND für andere Verfahren (ZestXML)

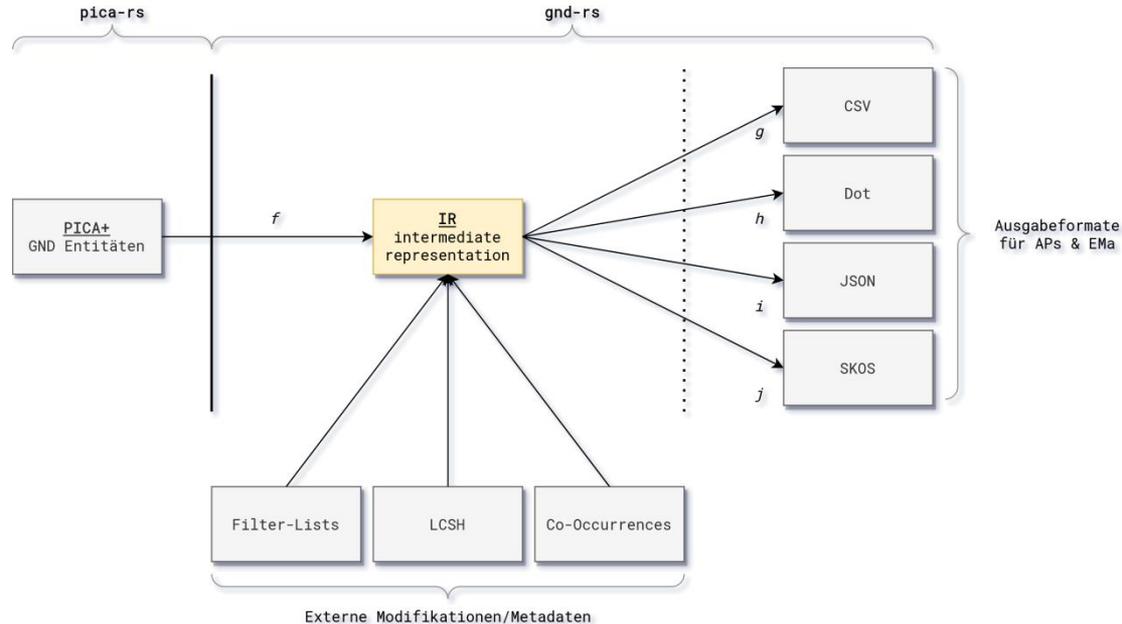
Herausforderungen

- Verschiedene Sichtweisen auf die GND
- ML-Verfahren erfordern eine gezielte Aufbereitung der Terminologie
- Eignung für maschinelle Verfahren, Aktualität und Vollständigkeit der GND
- Manipulation der Terminologie

Anforderungen

- schnelles, iteratives Arbeiten
- reproduzierbare Manipulationen an der GND
- schnelles Hinzufügen experimenteller Features
- unterschiedliche Ausgabeformate
- Integration in DVC-Pipelines

Architekturentwurf



Demo: Skosifizierung der GND

```
$ gnd --config contrib/AEN.toml tests/data/118515551.dat.gz
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX gnd: <http://d-nb.info/gnd/>
```

```
gnd:118515551 a skos:Concept;
  skos:altLabel "브로흐, 헤르만"@de,
    "Broxi, Herman"@de,
    "Broch, German"@de,
    "ברוך, הרמן"@de,
    "ヘルマン・ブロッホ"@de,
    "Broch, Herman"@de;
  skos:hiddenLabel "הרמן ברוך"@de,
    "German Broch"@de,
    "Herman Broxi"@de,
    "Herman Broh"@de,
    "헤르만 브로흐"@de;
  skos:prefLabel "Broch, Hermann (1886-1951)"@de.
```

Demo: Tabellieren der GND

```
$ gnd --config contrib/AEN.toml tabulate synonyms tests/data/118515551.dat.gz
uri,kind,synonym
http://d-nb.info/gnd/118515551,alternative,"Mproch, Cherman,,
http://d-nb.info/gnd/118515551,alternative,"뵁-뵁뵁뵁, 쵁 쵁뵁뵁 뵁뵁"
http://d-nb.info/gnd/118515551,hidden,"Broch, Hermann"
http://d-nb.info/gnd/118515551,hidden,German Broch
http://d-nb.info/gnd/118515551,alternative,"Broxi, Herman"
http://d-nb.info/gnd/118515551,alternative,"Broch, Herman"
http://d-nb.info/gnd/118515551,preferred,"Broch, Hermann (1886-1951)"
http://d-nb.info/gnd/118515551,hidden,Herman Broxi
http://d-nb.info/gnd/118515551,alternative,"Broch, German"
http://d-nb.info/gnd/118515551,hidden,"쵁 쵁뵁뵁 뵁뵁 뵁-뵁뵁뵁"
...
```

Demo: Konfigurationsdatei

```
[concept]
filter = "002@.0 =~ '^T[bfgpsu][1z]$\ ' && 008A.a == 's' && !008@a? && 007K.a == 'gnd'"
skip_invalid = true
min_synonym_length = 4
synonym_filter = "\\(.*Motiv\\)"$
base_uri = "http://d-nb.info/gnd/"
translit = "nfd"
```

```
[[collection]]
name = "GND Entity Codes"
path = "004B.a"
base_uri = "http://d-nb.info/aen/entity-code#"
filter = "^[a-z]+$"
```

```
[skosify]
language_tag = "de"
pretty = true
```

Ausblick

- Anreicherung der GND mit Synonymen aus externen Thesauri (LCSH)
- Verwendung von Teil-Terminologien im Ensemble
- Filtern von Synonymen
- Potentiale der GND besser nutzen (Relationen, Entitäten-Codes)

Links

Der Quellcode zu den Tools *gnd* und *pica* findet sich auf *GitHub*:

- <https://github.com/deutsche-nationalbibliothek/gnd-rs>
- <https://github.com/deutsche-nationalbibliothek/pica-rs>

Danke für Ihre Aufmerksamkeit.

n.wagner@dnb.de