Machine-based subject indexing and beyond for scholarly literature in psychology at ZPID

A case study of how we use Annif

Florian Grässle, Tina Trillitzsch
Leibniz Institute for Psychology (ZPID)

Semantic Web in Libraries 2023 (SWIB23), Sept. 13, 2023

SWIB23
Semantic Web in Libraries

Intro & Outline

Part 1: **Context**

- Introducing our reference database PSYNDEX
- How we index, our controlled vocabularies

Part 2: **Automatic Indexing in PSYNDEX**

- past: old system AUTINDEX,
- now: our Annif setup,
- future plans with Annif

Part 3: **Annif Special – Improving performance, using it beyond "just" subjects**

- what we do when some terms are suggested **not often enough** or suggested **too often**
- marking some subject suggestions as main topics
- using Annif for other, study-level vocabularies

ZPID and PSYNDEX

ZPID (Leibniz Institute for Psychology):

- **publicly funded** Open Science institute **for psychology** in German-speaking countries
- supports **researchers**, practicing psychologists, students & professors, but also lay people

PSYNDEX – most well known service:

- **reference database of psychological literature** – mostly **scholarly articles** and books, plus selected popular science materials
- **English and German** publications
- **1,000** new publication records/month

Indexing in PSYNDEX, controlled vocabularies

Several psychology-specific controlled vocabularies:

- **PSYNDEX "Controlled Terms"** ("CT") - largest, more on its own slide!
- **Subject Heading** Classification ("SH", 157 terms, based on APA classification)
- **Controlled Methods** ("CM", 58 terms): mix of publication **genre**/type (handbook, conference proceedings) and **study type** (meta analysis, experimental study)

Information about a study's sample population:

- **Age Group** Classification ("AGE", 12 terms)
- **Population Location** – country or continent ("PLOC")

We want Annif to help us with all of these, if possible! Currently, PSYNDEX Controlled Terms in production. SH is ready, rest: in testing.

Our Skosmos instance is (largely) public now! Check out https://vocabs.leibniz-psychology.org

Vocabulary: PSYNDEX Controlled Terms ("CT")

- Most important, largest vocabulary: over **6,800 terms**, complex hierarchy
- **English and German** labels and synonyms
- Based on American Psychological Association's **APA Thesaurus of Psychological Index Terms**, German translation by us;
- **licensed from APA** (limits open availability)
- **Biannual updates** from APA; we translate the new additions
- Indexers use "**weighting**" – marking some applied controlled terms as **main topics** (can be used for ordering search results by relevance – publication is more relevant if it includes searched term as "weighted")

Automatic Indexing in PSYNDEX: Then and now

- For 15 years (from 2006): **lexical system**, AUTINDEX, for suggesting terms from **Controlled Terms**
- It processed documents **overnight**, writing suggestions into **special fields** in our database, then displayed in the UI to be copied/used
- Today, AUTINDEX is unstable, unmaintained and **outdated**.
- Since February 2023, Controlled Terms are suggested by ***Annif*** instead (developed at National Library of Finland)

On the following slides:

- how we **switched** from AUTINDEX to **Annif**
- our **Annif setup**: corpus generation and splitting, ensemble and backends, optimizations, etc
- ambitious **plans for the future** with Annif: even more vocabularies, fully automatic indexing for a segment of our records and what that will require

From AUTINDEX to Annif

Currently, Annif is a **drop-in replacement** for AUTINDEX:

- **suggestions** for **Controlled Terms** only
- written into same record field AUTINDEX wrote to (CTAI)
- based on an Annif **ensemble** – combination of
  - **machine learning** trained on human-indexed documents
  - **lexical components** (similar to old system AUTINDEX)

How we arrived there:

Corpus generation and splitting: Based on year, document type, parts used for training and testing/optimization. See next slide.

Backend selection: **omikuji-bonsai** (machine learning) + **MLLM & STWFSA** (lexical, to help with new, not yet learned additions to vocabulary)

Optimization: Annif's `hyperopt` command to determine how much **each backend should contribute** to combined **ensemble** for best performance (testing 200 random ratios)

AUTINDEX and Annif in our database

```
<CTAI>|e Drawing |d Zeichnen</CTAI>
<CTAI>|e Cues |d Hinweisreize</CTAI>
<CTAI>|e Measurement |d Messung</CTAI>
<CTAI>|e Models |d Modelle</CTAI>
<CTAI>|e Roles |d Rollen</CTAI>
<CTAI>|e Simulation |d Simulation (Methodik)</CTAI>
<CTAI>|e Testing |d Testen</CTAI>
<CT>|e Drawing |d Zeichnen |g x</CT>
<CT>|e Spatial Imagery |d Räumliche Bildvorstellung |g x</CT>
<CT>|e Spatial Organization |d Räumliche Organisation (Wahrnehmung)</CT>
<CT>|e Cues |d Hinweisreize</CT>
<CT>|e Cognitive Ability |d Kognitive Fähigkeiten |g x</CT>
<CT>|e Childhood Development |d Entwicklung in der Kindheit |g x</CT>
<CT>|e Cognitive Flexibility |d Kognitive Flexibilität |g x</CT>
```

CTAI - Controlled Terms suggested by AUTINDEX or Annif

CT - Controlled Terms finally chosen by human indexer

XML record of a publication, showing term fields CTAI and CT

Now: Annif Corpus Preparation & Splitting

All **400,000** records (title, abstract, author-supplied uncontrolled keywords)

Only keep:

- published **2014 or later** (determined after much optimization and testing – as current as possible to train new additions to vocabulary)
- **not a "Collection"** of papers, like conference proceedings (indexed only very broadly, not representative)

German documents

- Training data: 80% of German records
- Test & Validate: 20%, only newest: >= 2021
  - 10% Test – check suggestions against human-indexed "gold standard"
  - 10% Validate (auto-optimizing parameters, backend ratios)

English documents

- Training: 80% of English records
- Test & Validate: 20%
  - 10% Test
  - 10% Validate

Future plans with Annif: Going fully automatic

Fully automatic indexing for *some* publication genres:

- human indexing (supported by Annif suggestions) for our "**core**" genre, **research papers**
- "**non-core**" publication genres could be auto-indexed: patient information, self-help books, textbooks, interviews …

Requirements:

- **classifying** documents as **"core" vs "non-core"** based on genre or study type – we have a **vocabulary** for that: **CM** (Controlled Methods)! Can Annif do the classification?
- teaching Annif to mark some of its suggestions (main topics) as "**weighted**"

⇒ We'll discuss both in the next part: "Annif Special", among other things.

Annif Special: Performance improvements; beyond subjects

To achieve our "future plans with Annif", we tried a few thing that may be interesting:

1. Improving Annif's performance:
    - When terms are **not suggested often enough** (high false negative rate)
    - When terms are suggested **too often** (high false positive rate)

2. Beyond "simple" subject indexing
    - Getting Annif to mark some of its suggested controlled terms as **"weighted"** (main topics)
    - Getting Annif to predict even more specific things that are not "subjects" per se:
        - **document genre/study type** (CM vocabulary)
        - **age group** and **location** of a study's population/sample

Term not (correctly) suggested often enough

Example reported by our indexers:
Covid-19

Reasons we found:
Not well (or at all) represented in corpus ⇒ **untrained** by machine-learning backend

- either because concept **new to the vocabulary**
- or not used much in the past

Solutions:
For new concepts: Fine-tune **corpus year** coverage: make **training set as recent** as possible. Not a complete fix - we'll **never catch up** with the updates – so…

Lexical backend components need to take over: Activate use of `skos:hiddenLabels` in MLLM component. Add more of these labels to problematic terms if needed.

Term suggested too often (incorrectly): Blocking

Examples reported by indexers:
Management, Health, Treatment, Learning, Behavior, Sex.

Reason: Terms are too broad/general to use in most cases.

Solution: Create **blocklists** – lists of terms that won't be suggested anymore (or less often/only by some backends), preferably create them **automatically** (always up-to -date with corpus and vocabulary)!

How we "block" in Annif: Creating a new "reduced" skos vocabulary where concepts to block are marked `owl:deprecated true` (Annif will never suggest those).

Small problem: Annif doesn't support separate vocabularies (=blocklists) per backend component, only whole ensemble ⇒ concepts will be *fully* blocked (never suggested at all anymore), not just by backend component that suggests them too often!

Temporary solution: only use block lists for fully automatic indexing; for "supported" human indexing, mark suggestions as "possibly too general", but still display them.

Term suggested too often (incorrectly): Blocking

Algorithm: Check subject-level "goodness" of each terms in vocab (calculated by Annif using `results-file` option of `annif eval` command).

- For terms with "bad" values (high false alarm rate > 0.1), calculate a **score**: Add 1 score point for each of several indicators that term is too broad.
- If score > threshold (for now, 2): add term to blocklist.
- **Human** indexers **judge blocklist candidates** (should definitely be blocked/not blocked). Use aggregated results to **fine-tune** algorithm **thresholds**. Repeat.

Indicators for too broad terms, adding to score:

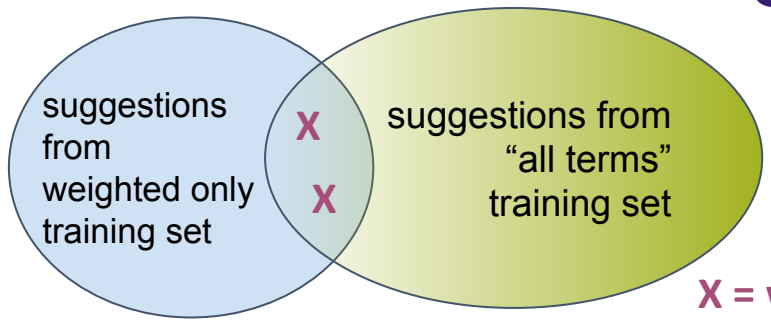
- if in a **low percentile (**under 40th) = not used by humans often, compared to least and most used concepts: **+ 1**
- is **top concept** in vocab hierarchy: **+1**
- is marked as "**conceptually broad, please use a more specific term**" in scope note by vocab editors: **+1**
- *also* has a **high false positive** rate (miss rate = rarely suggested correctly): **+1**

Getting Annif to suggest some terms as "weighted"

Weighted terms: up to 5 **main topics** of a document

To automatically mark some of Annif suggestions as main topics:

- we created a **separate training corpus** training only the subset of weighted terms
- plus regular corpus trained on all terms
- terms that Annif suggests for the document using ***both*** training sets (**intersection**) are marked weighted



X = weighted terms

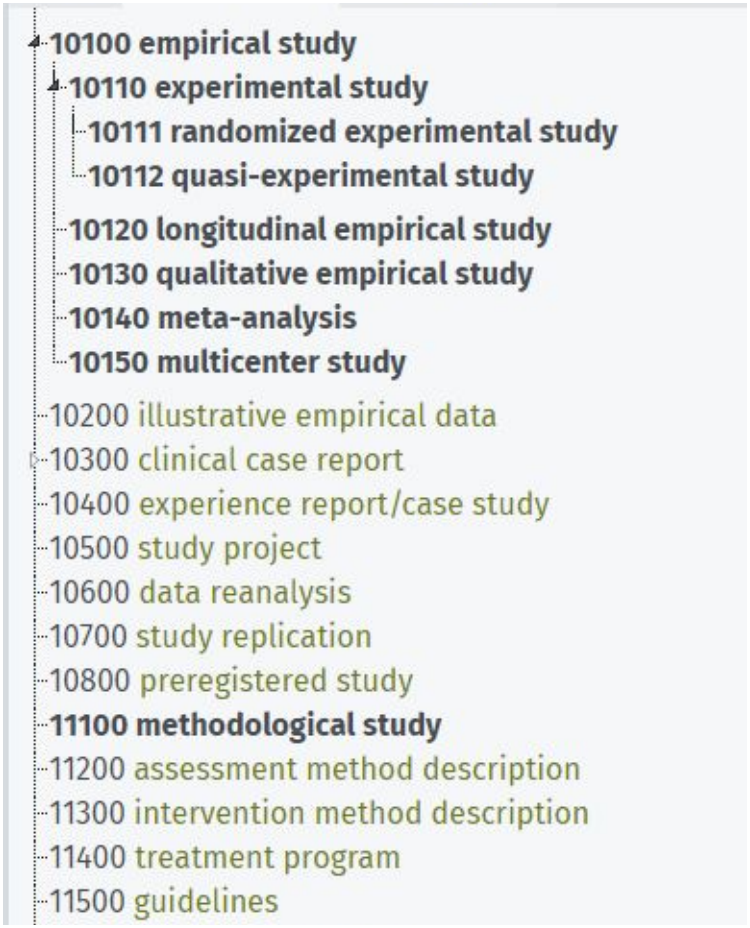
All terms suggestions	Notation	Score
Cognitive Flexibility	65259	0.9484
Drawing	15050	0.4806
Spatial Perception	48900	0.2361
Childhood Development	08760	0.2337
Cognitive Development	10080	0.1402
Cues	12680	0.1232
Weighted terms suggestions		
Cognitive Flexibility	65259	0.6636
Drawing	15050	0.2748
Childhood Development	08760	0.1573
Cognitive Development	10080	0.1313
Learning	28030	0.0494

Other Vocabularies: CM

Controlled Methods (CM) vocabulary: used in PSYNDEX to describe publication genre *and* study method used in scholarly articles.

We want Annif to predict CMs:

- for their own sake – to annotate our documents with them,
- but also to **classify** documents as core or non-core **based on predicted genre**, to tell us which need **human indexing** ("empirical study" and subconcepts, and "methodological study") or can be **auto-indexed** (anything else)



Other Vocabularies: CM

Where we are:

- We trained Annif on our corpus with CMs
- First results: promising performance compared to human-indexed gold standard

Todo:

- Further testing, comparing to other, non-Annif ways of classifying into two "buckets" (e.g. one-shot learning).
- If we end up using Annif: splitting into two groups based on predicted CM-type, one sent to auto-indexing, one to generate suggestions sent to human indexers

Other Vocabularies: AGE, Pop. location

Both: study-level – describe sample population. Suggestions need to be specific and true!

Age Group: Very small vocabulary (see screenshot →).

Population Location: A few hundred country names.

Where we are: AGE is trained and performs surprisingly well – but: better at suggesting **top concepts** than their subconcepts. Caused by artificial "up-posting" in the past (e.g. "Childhood" always automatically added when "Infancy" was used).

Todos: AGE – Fix/remove "up-posting" if possible, else only suggest top concepts. Pop. Location: Generate vocabulary of locations, train and test.

Workflow now

New documents arrive

↓

annif
suggests <CT>

↓

🧑 Human indexer
chooses <CT>
adds <AGE> <CM> <SH>
adds prereg links & more

Manual indexing; controlled terms are based on **suggestions by Annif**

… & in the future

New documents arrive

↓

annif classifies core vs non-core

Core documents

↓

annif
suggests <CT><AGE><CM><SH>

↓

🧑 Human indexer
chooses <CT><AGE><CM><SH>
adds prereg links & more

Manual indexing; terms, age, methods and subject headings are based on **suggestions by Annif**

Non-core documents

↓

annif
chooses <CT><AGE><CM><SH>

Controlled terms are trained on **reduced vocab** (terms on blocklist are excluded)
Controlled terms intersected with suggested **weighted terms** are marked "weighted"

Fully automatic, unsupervised indexing by Annif for terms, age, methods and subject headings

Summary

Part 1: **PSYNDEX, how we index, vocabularies**

Part 2: **Automatic Indexing in PSYNDEX**

- how we replaced our old system with Annif; our setup
- future plans: full automation for part of our records and what that requires

Part 3: **Annif Special**

- Improving performance (too seldom, too often)
- Beyond subject indexing (weighting, genre/study type, study population)

Thank you for listening!

Questions, comments?

ZPID vocabs browser (Skosmos): https://vocabs.leibniz-psychology.org

PSYNDEX: https://psyndex.de