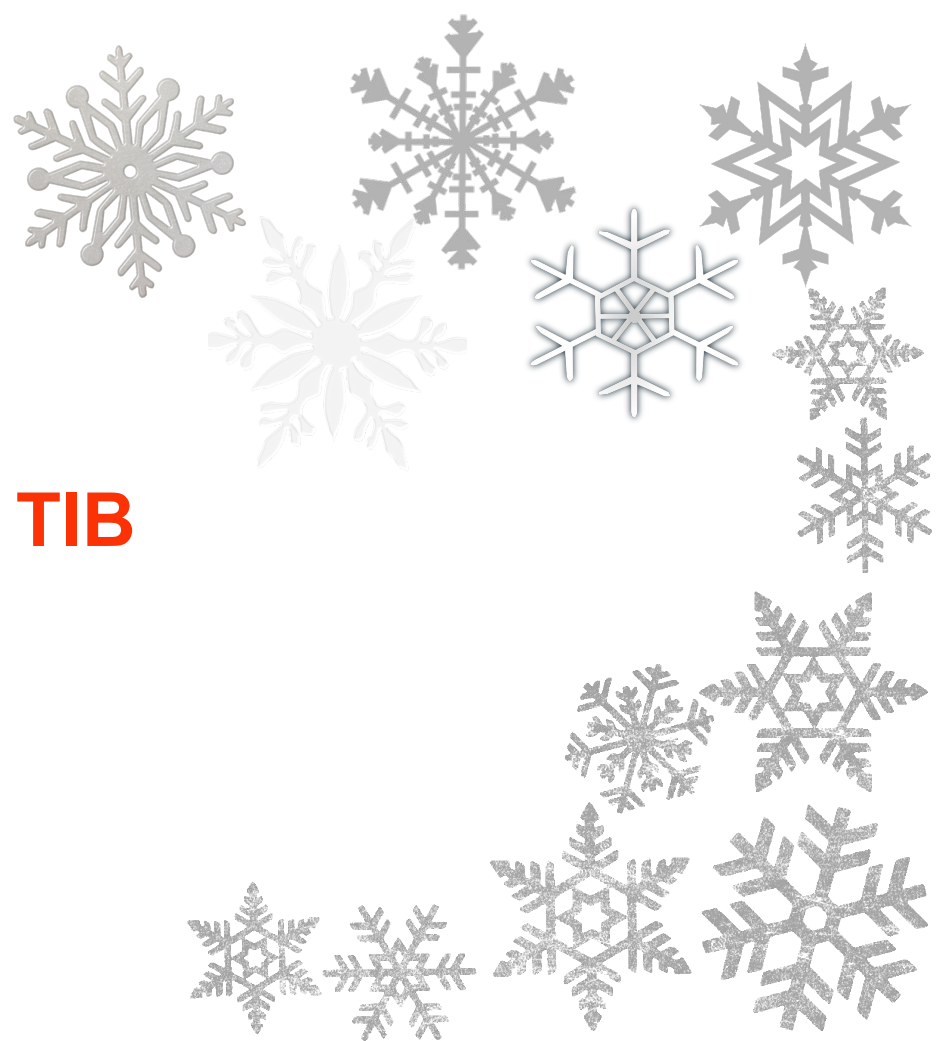


TIB

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK



annif an der TIB

Susanne Arndt, Berrit Genat, Mila Runnwerth
Austausch innerhalb der AG SE@TU9
11. Dezember 2020

Agenda



Kurzvorstellung von annif

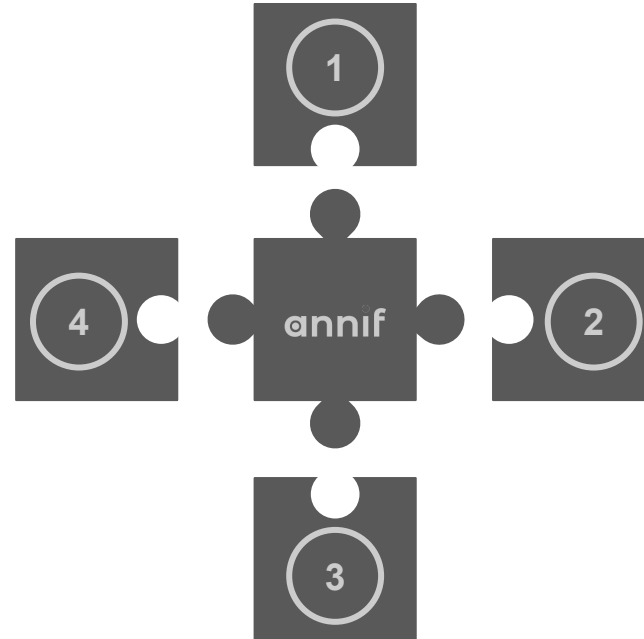
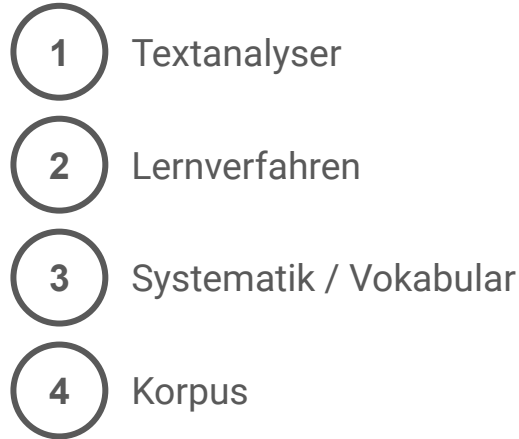
Datenaufbereitung für den Einsatz von annif

Anwendungsbeispiel LinSearch

Kurzvorstellung von annif



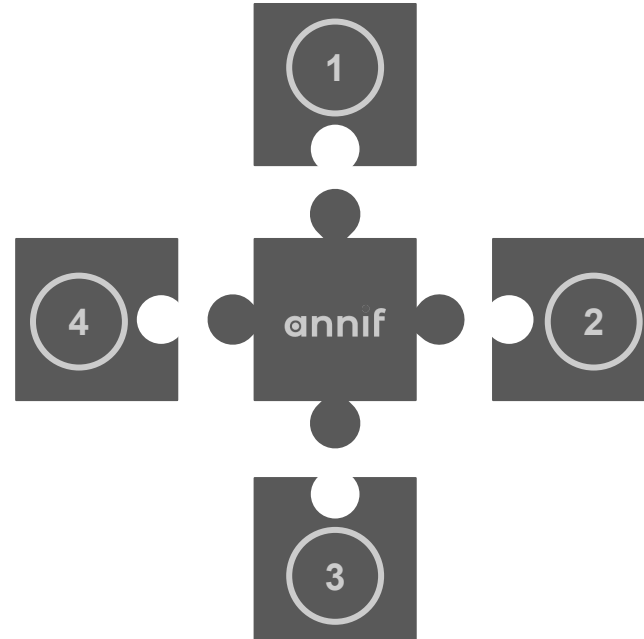
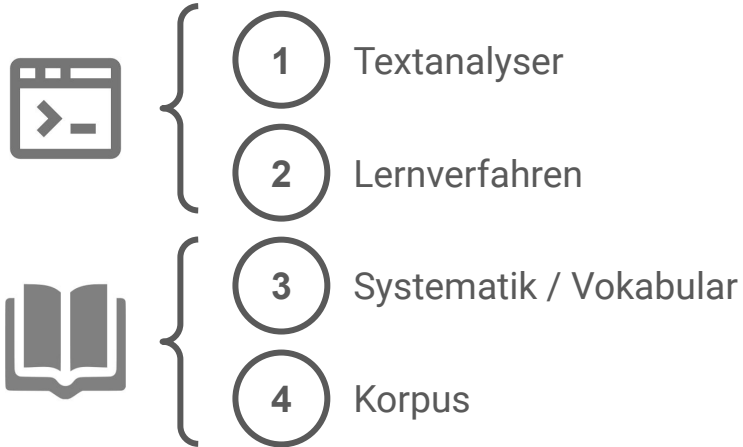
- Entwickelt an der Finnischen Nationalbibliothek
- Projektseite: <https://annif.org/>



Kurzvorstellung von annif



- Entwickelt an der Finnischen Nationalbibliothek
- Projektseite: <https://annif.org/>



Kurzvorstellung von annif



1 Textanalyser

Simple Analyzer

Snowball Analyzer

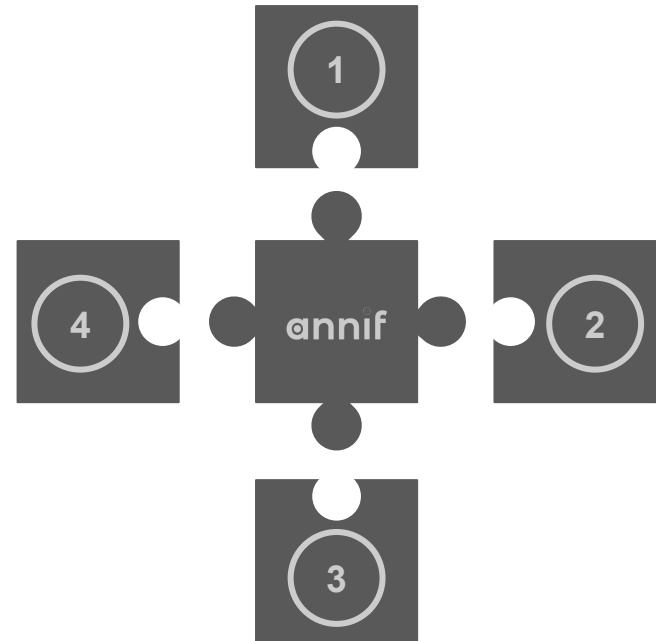
2 Lernverfahren

Alnum Analyzer

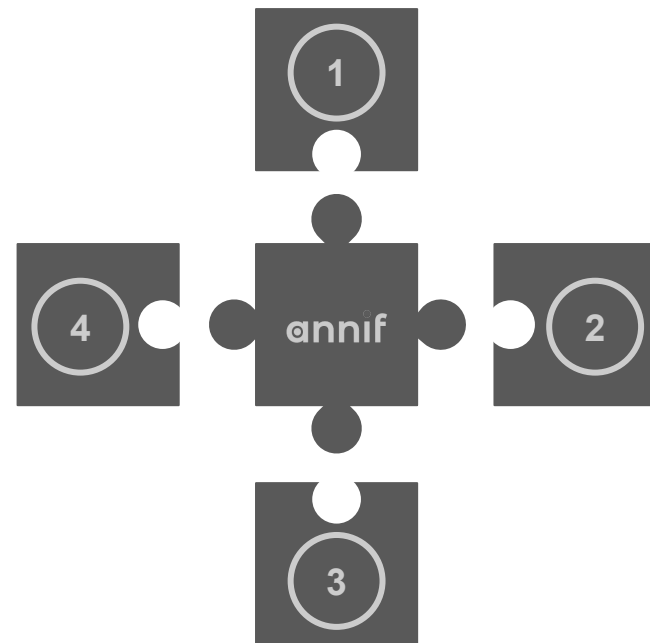
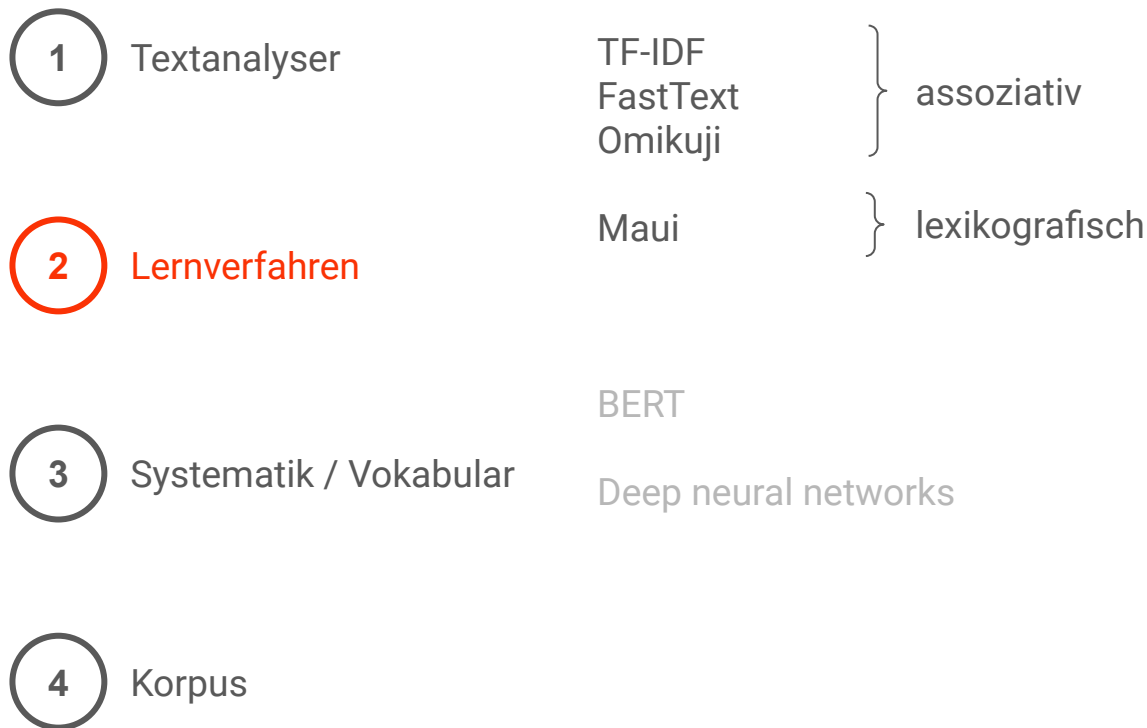
CJK Analyzer

3 Systematik / Vokabular

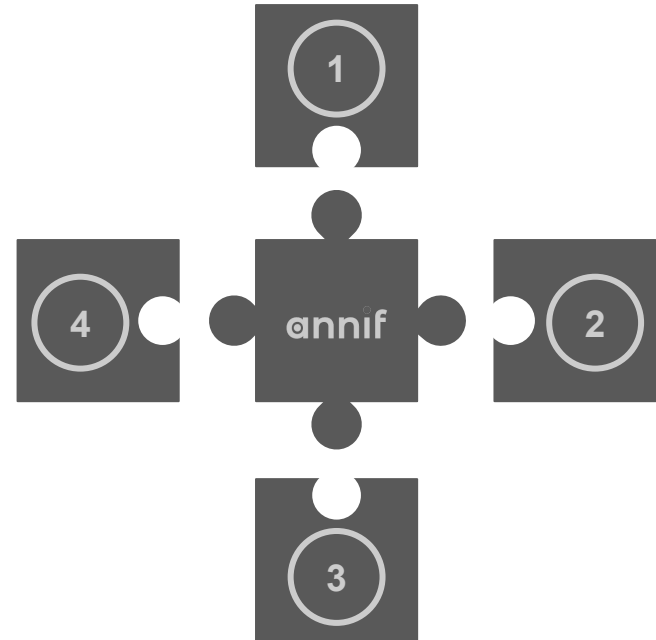
4 Korpus



Kurzvorstellung von annif



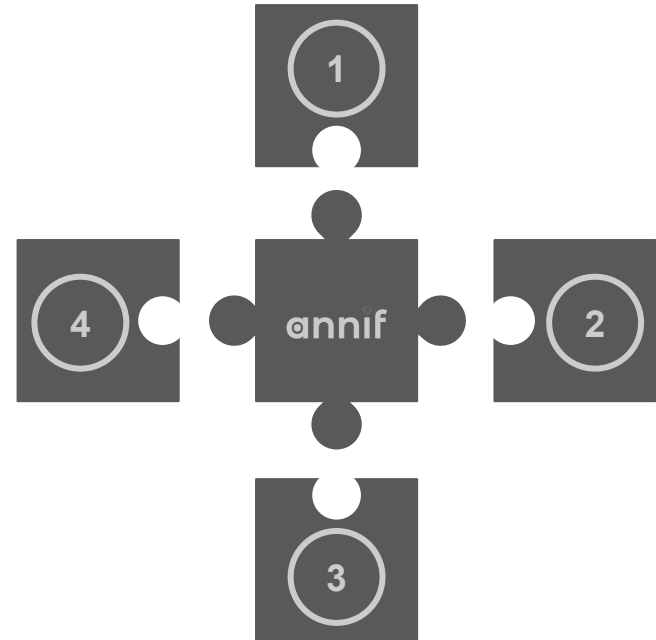
Kurzvorstellung von annif



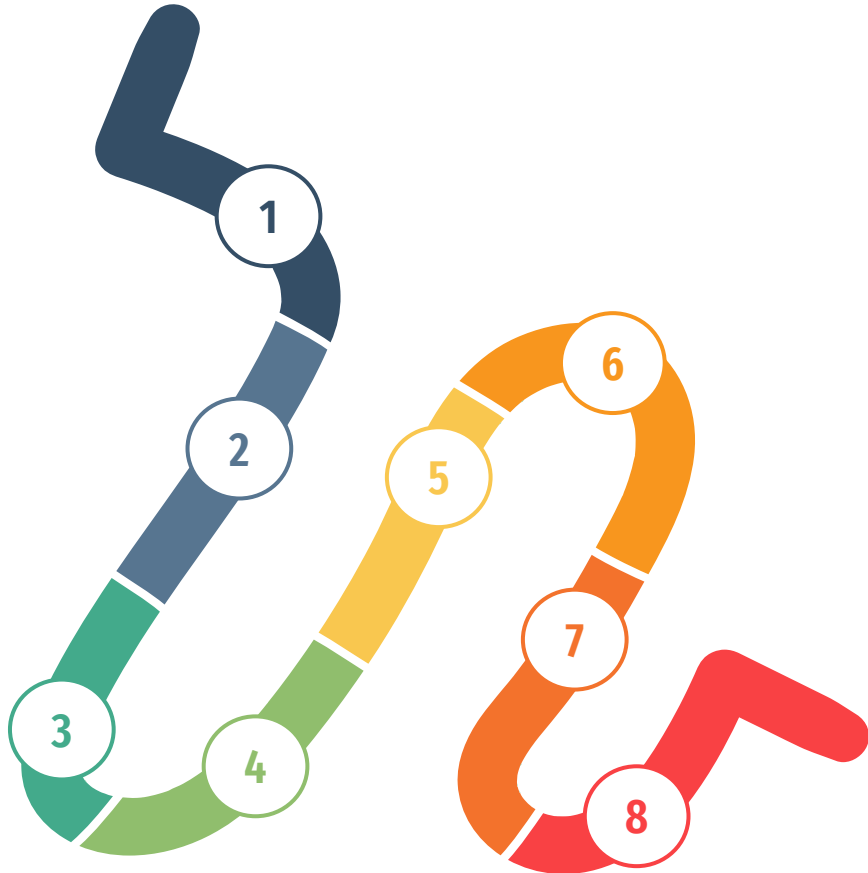
Kurzvorstellung von annif



- 1 Textanalyser
TIB-Index
FIDmove-Index
- 2 Lernverfahren
Ihr Index
- 3 Systematik / Vokabular
- 4 Korpus



Datenaufbereitung: Der Weg ist das Ziel



- | | | | |
|---|---------------------------------|---|-------------------------|
| 1 | Vokabular transformieren | 5 | Goldstandard definieren |
| 2 | Datenkollektion zusammenstellen | 6 | Testmenge extrahieren |
| 3 | Datendump herunterladen | 7 | annif konfigurieren |
| 4 | Daten transformieren | 8 | Ergebnisse analysieren |

1

Vokabular transformieren



Dauer (A): 1 min bis 1 Tag

Dauer (B): 1 h bis 1 Tag



https://github.com/runwertb/annif_automated_indexing_at_tib/wiki/Vokabular-transformieren



	Variante A: Lookup-Table (.tsv)	Variante B: SKOS-Datei (.ttl)
Ziel	2 Spalten <ul style="list-style-type: none">• Spalte 1: 1 URI pro Klasse• Spalte 2: Bezeichner der Klasse	SKOS-Modellierung des eigenen Vokabulars
Input	HTML, Tabelle (Idealfall), XML, RDF	HTML, Tabelle (Idealfall), XML, RDF (nicht in SKOS)
Tool	z.B. OpenRefine, Protégé	z.B. OpenRefine, Protégé
Teilaufgaben	<ul style="list-style-type: none">• ggf. URI pro Vokabularterm erstellen• Infos aus einer RDF-Modellierung raussparqln	<ul style="list-style-type: none">• eigene Daten durch die SKOS-Brille sehen• ggf. URI pro Vokabularterm erstellen• ggf. Beziehungen explizit machen• Tripel anlegen• Turtle-Datei anlegen
Voraussetzungen	GREL oder Python	GREL oder Python, Basisverständnis über SKOS-Elemente und RDF-Tripel

1

Vokabular transformieren



Input

	A	B	C	D	
1	Systemstelle		Link	Beschreibung	Hi
2	mat 1		x	Allgemeines	
3	mat 1	mat 1.4	x	Didaktik der Mathematik. Programmierer und rechnergestützter Mathematikunterricht	
4	mat 1	mat 1.5	x	Beruf. Ausbildung. Nachwuchs	

https://github.com/runwertb/annif_automated_indexing_at_tib/wiki/Vokabular-transformieren



1

Vokabular transformieren



https://github.com/runwertb/annif_automated_indexing_at_tib/wiki/Vokabular-transformieren



Variante A - Output

```
1 URI Beschreibung
2 <http://unicorn.fly/tib_lok_sys#mat1> Allgemeines
3 <http://unicorn.fly/tib_lok_sys#mat1_4> Didaktik der Mathematik.
  Programmierer und rechnergestützter Mathematikunterricht
4 <http://unicorn.fly/tib_lok_sys#mat1_5> Beruf. Ausbildung. Nachwuchs
```

Variante B - Output

```
1 @prefix skos: <http://www.w3.org/2004/02/skos/core#> .
2 @base <http://unicorn.fly/tib_lok_sy> .
3 @prefix math: <http://unicorn.fly/tib_lok_sy#> .
4
5 math: a owl:Ontology ; owl:imports skos: .
6
7 <http://unicorn.fly/tib_lok_sys#mat1> skos:prefLabel "Allgemeines"@de .
8 <http://unicorn.fly/tib_lok_sys#mat1> a skos:Concept .
9 <http://unicorn.fly/tib_lok_sys#mat1_4> skos:prefLabel "Didaktik der Mathematik. Programmierer und
  rechnergestützter Mathematikunterricht"@de .
10 <http://unicorn.fly/tib_lok_sys#mat1_4> a skos:Concept .
11 <http://unicorn.fly/tib_lok_sys#mat1> skos:narrower <http://unicorn.fly/tib_lok_sys#mat1_4> .
12 <http://unicorn.fly/tib_lok_sys#mat1_5> skos:prefLabel "Beruf. Ausbildung. Nachwuchs"@de .
13 <http://unicorn.fly/tib_lok_sys#mat1_5> a skos:Concept .
14 <http://unicorn.fly/tib_lok_sys#mat1> skos:narrower <http://unicorn.fly/tib_lok_sys#mat1_5> .
```

2

Datenkollektion zusammenstellen



Homepage

OAI-Schnittstelle

Die Metadaten zum Abruf über OAI-PMH

Über die OAI-Schnittstelle der TIB können Metadaten in den Formaten OAI Dublin Core, MARC XML oder RDF XML heruntergeladen werden, eine tägliche Aktualisierung ist möglich. Details siehe bitte in der [Schnittstellen-Dokumentation](#). Es gibt verschiedene Tools, die das Herunterladen unserer Daten über OAI-PMH ermöglichen. Ein Tool mit graphischer Oberfläche ist beispielsweise [MarcEdit](#).

TIBKAT-Daten

Beispiel-Links (Komplettabzug, tagesakuell)

- Dublin Core: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=oai_dc&set=tibkat
- MARC XML: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=marc_xml&set=tibkat
- RDF XML: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=rdf_xml&set=tibkat
- RDF JSON: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=rdf_json&set=tibkat

Metadaten der frei verfügbaren elektronischen Bestände

Beispiel-Links (komplette Daten, tagesakuell)

- Dublin Core: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=oai_dc
- MARC XML: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=marc_xml
- RDF XML: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=rdf_xml
- RDF JSON: https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=rdf_json

Metadaten des TIB AV-Portals

Beispiel-Links (tagesakuell)

- Dublin Core: https://www.tib.eu/oai/public/repository/av-portal?verb=ListRecords&metadataPrefix=oai_dc
- MARC XML: https://www.tib.eu/oai/public/repository/av-portal?verb=ListRecords&metadataPrefix=marc_xml
- RDF XML: https://www.tib.eu/oai/public/repository/av-portal?verb=ListRecords&metadataPrefix=rdf_xml

Die Metadaten als Gesamtabzug (Dump)

TIBKAT-Daten und Metadaten der frei verfügbaren elektronischen Bestände

Die Katalogdaten der TIB und die Metadaten der frei verfügbaren elektronischen Bestände sind komprimiert als Gesamtabzug in den Formaten JSON-LD, N3 und Turtle abrufbar:

- JSON-LD: https://tib.eu/data/roff/open_isoid.dump.gz
- N3: https://tib.eu/data/roff/open_n3.dump.gz
- Turtle: https://tib.eu/data/roff/open_turtle.dump.gz

RDF-Export von Metadaten des TIB AV-Portals

Die autoritativen sowie zeitbasierten, automatisch generierten Metadaten von Filmen des AV-Portals der TIB sind ebenfalls komprimiert als Gesamtabzug in den Formaten JSON-LD, N3 und Turtle abrufbar. Diese Datensätze finden Sie auf der [Download-Seite des AV-Portals der TIB](#).

<https://www.tib.eu/de/die-tib/open-data>

OAI 2.0 Request Results

<https://oai.library.tib.eu/oai20/open?verb=ListRecords>

There are errors in the request: Error: Invalid request: The verb 'ListRecords' is not supported. Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

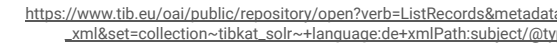
Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.

Response: Error: Invalid request: The verb 'ListRecords' is not supported.



https://www.tib.eu/oai/public/repository/open?verb=ListRecords&metadataPrefix=marc_xml&set=collection~tibkat_solr~+language=de+xmlPath:subject/@type=gn

https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Definition-des-Daten-Dumps



2

Datenkollektion zusammenstellen



Datenquelle

- Alle Datenquellen
- TIBKAT (968.836)
- Online Contents (21.600)

Sprache

- Alle Sprachen
- Deutsch (968.836)
- Englisch (582.543)
- Französisch (30.282)
- Ohne Sprachangabe (30.028)
- Russisch (13.567)
- Spanisch (7.580)
- Italienisch (7.504)



https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Definition-des-Daten-Dumps



Dauer: ~ 10 Minuten

Portal

OAI-Schnittstelle

```
language%5D%5B1%5D=de
supplierPrefix%5D%5B0%5D=tibkat
locationCode%3A%28L%20mat%2A%20OR%20LB%20mat%2A%29
```

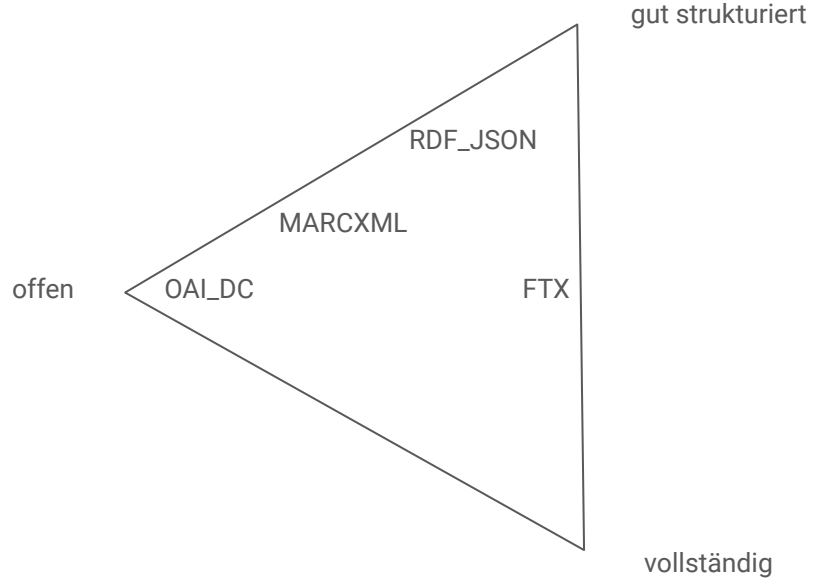
```
set=collection~tibkat
language:de
locationCode%3A%28L%20mat%2A%20OR%20LB%20mat%2A%29
```

2

Datenkollektion
zusammenstellen



Metadatenformat



https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Definition-des-Daten-Dumps



Datendump herunterladen



- Chunks mit 100 Metadatenätzen
- je nach Metadatenformat zwischen 2 - 20 Sekunden pro Chunk
- größter Korpus lag bei etwa 1.270.000 Datensätzen



Dauer: 5 - 50 Stunden

```
baseUrl="https://www.tib.eu/oai/public/repository/open?verb=listRecords&metadataPrefix=marc_xml&set=collection:tibkat_solr+language:de+xmlPath:subject/@type=end"
```

```
set -e
```

```
function token {  
  xmlstarlet sel -N xmlns="http://www.openarchives.org/OAI/2.0/" -t -v  
  '//_:OAI-PMH/_:ListRecords/_:resumptionToken[text()]' -nl $file  
}
```

```
i=0
```

```
token="__"
```

```
while [ -n "$token" ]; do  
  echo -n "$i "  
  file="oai-`printf %08d $i`.xml"  
  if [ $i -eq 0 ]; then  
    url=$baseUrl  
  else  
    url="$baseUrl&resumptionToken=$token"  
  fi  
  duration=$( { /usr/bin/time -f "%e" curl -s $url > $file; } 2>&1)  
  token=$(token $file)  
  echo $duration $url
```

```
i=$((i+1))
```

```
done
```

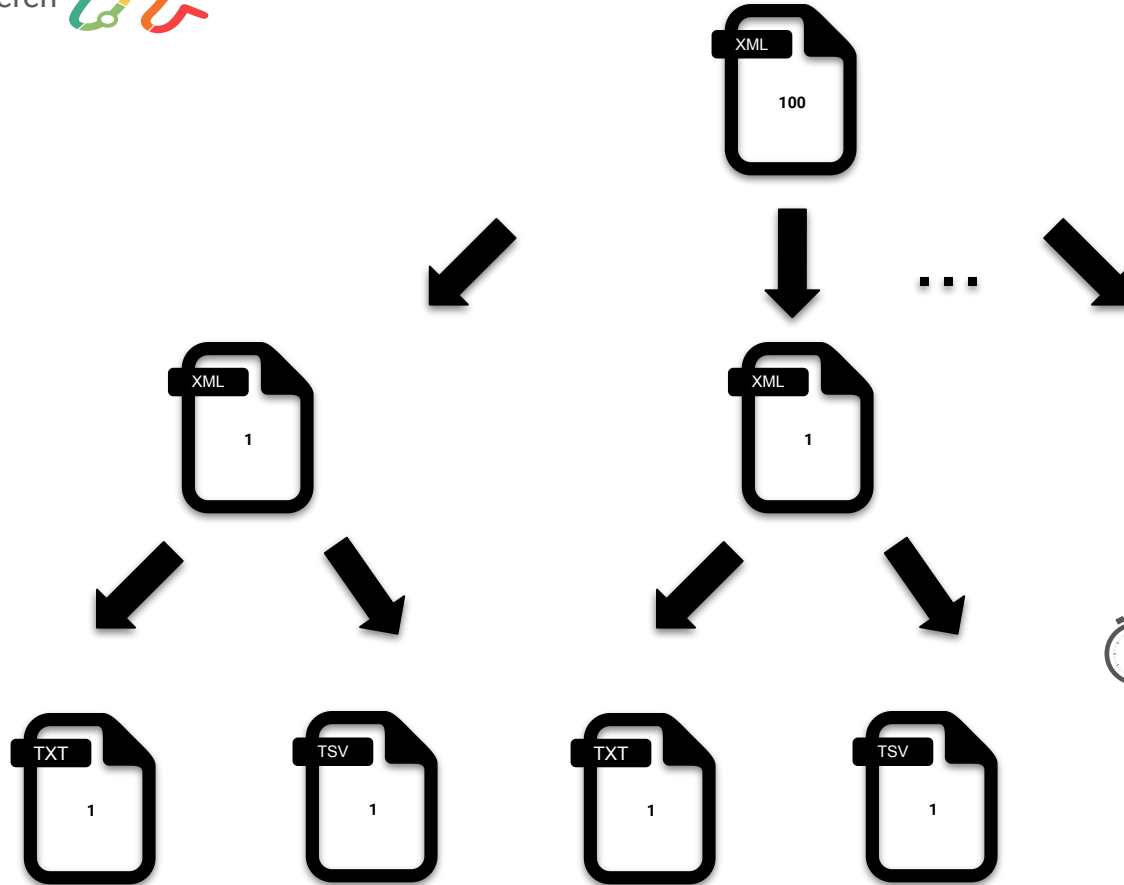


4

Daten transformieren



https://github.com/rumwerth/annif_automated_indexing_at_tib/wiki/%C3%9Cbersetzen-des-Dumps-in-die-annif-Syntax



Dauer: 3 - 15 Stunden



Daten transformieren



Der fertige Input für annif sieht dann pro Titeldatensatz so aus:

TXT: (Semantisch verwertbare) Metadaten

```
Führer auf den deutschen Schifffahrtstraßen  
Preußen, Ministerium der Öffentlichen Arbeiten  
Wasserbau  
Schifffahrt  
Deutschland
```

TSV: Ausgewählte Erschließungsdaten

```
<http://uri.gbv.de/terminology/bk/55.86> Schiffsverkehr, Schifffahrt  
<http://uri.gbv.de/terminology/bk/56.30> Wasserbau  
<http://uri.gbv.de/terminology/bk/15.40> Deutsche Geschichte: Allgemeines
```





MARCXML



```
xmlstarlet sel -N marcxml="http://www.loc.gov/MARC21/slim" -t -v  
'//marcxml:subfield[@code="2"][text()="bk"]/following-sibling::marcxml:subfield[@code="a"]  
' -nl tibkat00113689.xml
```

```
<marcxml:datafield ind1=" " ind2=" " tag="084">  
  <marcxml:subfield code="2">bk</marcxml:subfield<  
  <marcxml:subfield code="a">55.80</marcxml:subfield<  
</marcxml:datafield<  
<marcxml:datafield ind1=" " ind2=" " tag="084">  
  <marcxml:subfield code="2">bk</marcxml:subfield<  
  <marcxml:subfield code="a">50.01</marcxml:subfield<  
</marcxml:datafield<  
<marcxml:datafield ind1=" " ind2=" " tag="084">  
  <marcxml:subfield code="2">ssg</marcxml:subfield<  
  <marcxml:subfield code="a">19,2</marcxml:subfield<  
</marcxml:datafield<  
<marcxml:datafield ind1=" " ind2=" " tag="084">  
  <marcxml:subfield code="2">rvk</marcxml:subfield<  
  <marcxml:subfield code="a">ZG 8930</marcxml:subfield<  
</marcxml:datafield<  
<marcxml:datafield ind1=" " ind2=" " tag="084">  
  <marcxml:subfield code="2">loc</marcxml:subfield<  
  <marcxml:subfield code="a">T183.G32</marcxml:subfield<  
</marcxml:datafield<
```





Daten transformieren



FTX



```
xmlstarlet sel -T -t -m "//classification[@classificationName='bk']" -v  
'code[text()]' -o $('t' -v 'entries/entry[text()]' -nl
```

```
<classificationInfo  
  <classifications  
    <classification classificationID="106410458"  
      classificationName="bk">  
        <code>42.13</code>  
        <entries>  
          <entry>Molekularbiologie</entry>  
        </entries>  
      </classification>  
    <classification classificationID="106410148"  
      classificationName="bk">  
        <code>42.01</code>  
        <entries>  
          <entry>Geschichte der Biologie</entry>  
        </entries>  
      </classification>  
  <subjects>  
    <subject id="4071722-7" type="gnd">  
      <dc:subject xml:lang="de">Gentechnologie</dc:subject>  
    </subject>  
    <subject type="mesh">  
      <dc:subject xml:lang="de">Eugenics</dc:subject>  
    </subject>  
  </subjects>  
</classificationInfo
```





Goldstandard definieren



Dauer: ~ 10 Stunden



https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Anlegen-eines-Goldstandards



- Intellektueller Aufwand:
 - Repräsentanz des Gesamtkorpus ermitteln,
 - Qualität von hochwertiger Erschließung definieren,
 - passende Datensätze finden.
- Empfohlen: 10 - 15% vom Gesamtkorpus
- Das bedeutete in unserem größten Korpus von etwa 1.270.000 Metadatensätzen: ~12.700 Metadatensätze für den Goldstandard
- Wir haben diesem Fall geschummelt und nur 130 Metadatensätze ausgesucht.

AUSLEIHEN & BESTELLEN RECHERCHIEREN & ENTDECKEN LERNEN & ARBEITEN PUBLIZIEREN & ARCHIVIEREN FORSCHUNG & ENTWICKLUNG DIE TIB

Suche nach Büchern, Aufsätzen und mehr

Nur im Bibliothekskatalog der TIB suchen Hier geht es zum klassischen Katalog of

Mathematische Methoden der Physik (Deutsch)

Karbach, Michael
2017
ISBN: [3110456656](#) [9783110456653](#)
Buch / Print
<http://www.degruyter.com> <http://zbmath.org/?q=an> <http://is-eb.info/11029074...>

*Das Werk führt in die mathematischen Konzepte und ihre Anwendungen ein und ist sowohl vorlesungsbegleitend als auch aufgrund der vielen Beispiele zum Selbststudium bestens geeignet.
*This work introduces classical mathematical concepts and their applications and is designed to parallel a lecture course. Numerous

Wie erhalte ich diesen Titel?

TIB vor Ort
6 Exemplare
davon bestellbar
davon vorrätbar

TIB-Dokumentlieferung

Preisinformation

Details

Titel: Mathematische Methoden der Physik
Autor / Urheber: **Karbach, Michael**
Erschienen in: [De Gruyter Studium](#)
Verlag: De Gruyter
Erscheinungsort: Berlin, Boston
Erscheinungsjahr: 2017
Format / Umfang: X, 285 Seiten
Anmerkungen: 24 cm x 17 cm
Diagramme
Literaturverzeichnis: Seite 277-278
ISBN: [3110456656](#) [9783110456653](#)
Medientyp: Buch
Format: Print
Sprache: Deutsch
Schlagwörter: [Mathematische Physik](#) [Analysis](#)
Klassifikation: BKL: [31.42](#) Funktionen mit einer komplexen Variablen / [33.06](#) Mathematische Methoden der Physik / [31.46](#) Funktionsanalyse
MSC: [30B05](#) / [00A08](#) / [00A79](#)
DDC: [530.15](#)
Datenquelle: [TIBGAL](#)
Exportieren:

Ähnliche Titel



6

Testmenge
extrahieren



- Empfohlen: 10% vom Gesamtkorpus
- Auch hier haben wir gemogelt und nur eine handvoll Datensätze isoliert, um sie im WebGUI zu testen.
- Für uns erst interessant, wenn die Ergebnisse in Richtung Anwendbarkeit deuten.



Dauer: ~ 5 Minuten



7

annif konfigurieren



```
[TIBKAT_LokSysMath_de_OP]
name=TIBKAT_LokSysMath_de_OP
language=de
backend=omikuji
analyzer=snowball(german)
vocab=TIBLokSysMath

# [FIDmove_BK_de_maii]
# name=FIDmove_BK_de_maii
# language=de
# backend=maui
# endpoint=http://mauiserver:8080/mauiserver/
# tagger=FIDmove_BK_de_maii
# vocab=bk
# limit=1000
```

Konfiguration:
projects.cfg



Ablauf:
myAnnifProject.sh

```
#!/bin/bash

annif_project="TIBKAT_LokSysMath_de_OP"

#Load local classification
echo "Start loading local classification"
annif loadvoc $annif_project vocab/TIBLokSysMath.tsv
echo "Local classification loaded"
#Train loaded vocabulary
echo "Start training"
annif train $annif_project training/TIBIndex/TIBKAT_LokSysMath_de/
echo "Local classification trained"

#Evaluate against gold standard
annif eval $annif_project goldstandard/TIBIndex/TIBKAT_LokSysMath_de/

#Run Web GUI
echo "Start Web GUI"
unicorn --bind 0.0.0.0:5000"annif:create_app()"
#annif run
```



Dauer: 5 Stunden





Ergebnisse
analysieren




Evaluierungsfeature von annif




FIDmove-Index (TIBKAT-Anteil),
Basisklassifikation, deutsch, Maui

Precision (doc avg):	0.6076112412177986
Recall (doc avg):	0.6586065573770491
F1 score (doc avg):	0.5950772360608426
True positives:	147
False positives:	119
False negatives:	81
Documents evaluated:	122



TIB-Index, TIBKAT,
Lokalsystematik, deutsch, Omikujii

Precision (doc avg):	0.08139534883720931
Recall (doc avg):	0.813953488372093
F1 score (doc avg):	0.14799154334038056
True positives:	105
False positives:	1185
False negatives:	2
Documents evaluated:	129





Ergebnisse analysieren



Stichproben mit dem WebGUI



annif Web UI

Welcome!

See the [Swagger documentation](#) for an interactive REST API specification.

Mathematik für Chemiker

Verf.Beschr.: Differentialgleichungen, Quantenmechanik, Wahrscheinlichkeitsrechnung - wie alle exakten Naturwissenschaften erfordert auch die Chemie mathematisches Handwerkszeug, um Prozesse und Phänomene zu untersuchen. Was angehende Chemiker von der Mathematik wissen müssen, bietet in bewährter Weise "Mathematik für Chemiker" in der siebten Auflage. Das notwendige mathematische Rüstzeug wird massgeschneidert fürs Studium vermittelt, anschaulich in der Darstellung und ohne komplizierte Beweis Ketten. Zahlreiche praktische Beispiele aus der Chemie wecken das Interesse an der Mathematik und stellen den Bezug zur fachlichen Anwendung her. Die leicht verständliche Form garantiert den sicheren Einstieg, im Aufgabenteil mit Lösungen lässt sich das erworbene Wissen selbstständig überprüfen. Weiterführende Themen machen das Buch zum wertvollen Begleiter bis zum Examen. Durchgehend aktualisiert und um ein neues Kapitel zu numerischen Verfahren erweitert - für die Grundvorlesung Mathematik ebenso wie bei Fragen und Problemen im weiteren Studium unentbehrlich

PROJECT (VOCABULARY AND LANGUAGE)

TIBKAT_LokSysMath_de_OP

MAX # OF SUGGESTIONS

10 15 20

Get suggestions

SUGGESTED SUBJECTS

- Allgemeines
- Numerische Mathematik
- Wahrscheinlichkeitstheorie und mathematische Statistik
- Geschichte der Mathematik
- Geometrie
- Stochastische Prozesse. Stochastische Analysis.
- Gruppentheorie
- Differentialgleichungen.
- Randwertaufgaben
- Graphentheorie
- Approximationstheorie. Asymptotische Analysis

https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Ergebnisse



- 1 Vokabular transformieren
- 2 Datenkollektion zusammenstellen
- 3 Datendump herunterladen
- 4 Daten transformieren
- 5 Goldstandard definieren
- 6 Testmenge extrahieren
- 7 annif konfigurieren
- 8 Ergebnisse analysieren



Best Case:



~ 18h

Worst Case:

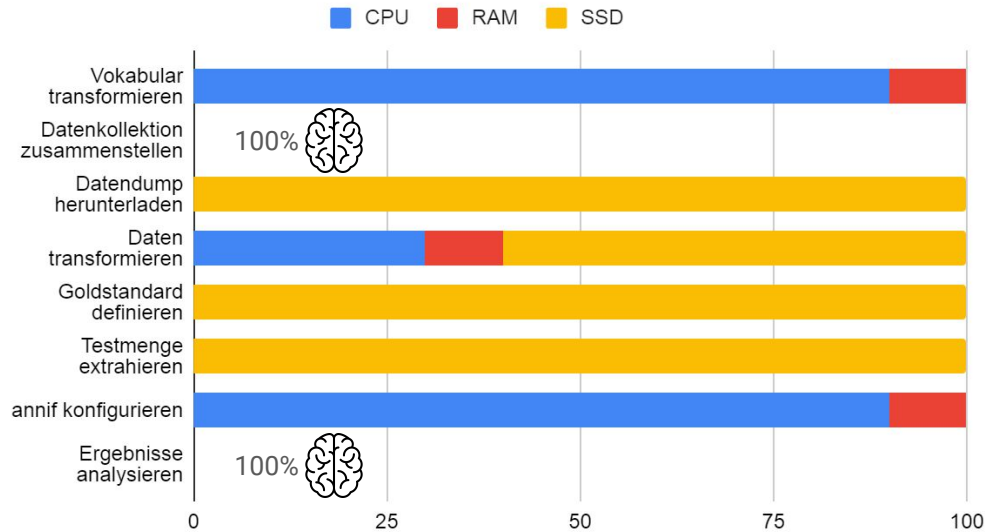


~ 90h

- 1
 Vokabular transformieren
- 3
 Datendump herunterladen
- 5
 Goldstandard definieren
- 7
 annif konfigurieren
- 2
 Datenkollektion zusammenstellen
- 4
 Daten transformieren
- 6
 Testmenge extrahieren
- 8
 Ergebnisse analysieren



CPU, RAM und SSD



Technische Mindestanforderungen (für 500.000 Metadatenätze)

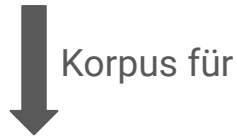
- CPU: Intel Core i7
- RAM: 16 GB
- SSD: M.2, 512 GB

Anwendungsbeispiel LinSearch



Automatische Fachzuordnung von Dokumentmetadaten.

Stufe 1*: Zuordnung nach vorhandenen Klassifikationselementen gemäß einer Konkordanz (Mapping)



Stufe 2: Automatische Fachzuordnung mit **annif**

1

Vokabular transformieren



```
<http://uricorn.fly/linsearch#tec> Technik allgemein
<http://uricorn.fly/linsearch#fer> Werkstoffkunde
<http://uricorn.fly/linsearch#mas> Maschinenbau, Energietechnik
<http://uricorn.fly/linsearch#elt> Elektrotechnik
<http://uricorn.fly/linsearch#ver> Verkehrstechnik, Verkehrswesen
<http://uricorn.fly/linsearch#bau> Bauwesen
<http://uricorn.fly/linsearch#ber> Bergbau
<http://uricorn.fly/linsearch#cet> Chemische und Umwelttechnik
<http://uricorn.fly/linsearch#med> Medizintechnik
<http://uricorn.fly/linsearch#arc> Architektur
<http://uricorn.fly/linsearch#che> Chemie
<http://uricorn.fly/linsearch#inf> Informatik
<http://uricorn.fly/linsearch#mat> Mathematik
<http://uricorn.fly/linsearch#phy> Physik
<http://uricorn.fly/linsearch#bio> Biowissenschaften, Biologie
<http://uricorn.fly/linsearch#pae> Erziehungswissenschaften, Fachdidaktiken
<http://uricorn.fly/linsearch#hor> Gartenbau
<http://uricorn.fly/linsearch#geo> Geowissenschaften, Geographie
<http://uricorn.fly/linsearch#his> Geschichte
<http://uricorn.fly/linsearch#lin> Sprachwissenschaften
<http://uricorn.fly/linsearch#lit> Literaturwissenschaften
<http://uricorn.fly/linsearch#phi> Philosophie
<http://uricorn.fly/linsearch#jur> Rechtswissenschaften
<http://uricorn.fly/linsearch#rel> Religionswissenschaft, Theologie
<http://uricorn.fly/linsearch#sow> Sozialwissenschaften
<http://uricorn.fly/linsearch#spo> Sportwissenschaft
<http://uricorn.fly/linsearch#oek> Wirtschaftswissenschaften
<http://uricorn.fly/linsearch#rest> Konnte nicht zugeordnet werden.
```



2

Datenkollektion zusammenstellen



Leichte Sprache Schnellzugriff English Anmelden

TIB LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

AUSLEIHEN & BESTELLEN RECHERCHIEREN & ENTDECKEN LERNEN & ARBEITEN PUBLIZIEREN & ARCHIVIEREN FORSCHUNG & ENTWICKLUNG DIE TIB

xmlPath:classification/@classificationProcedure=mapping

Nur im Bibliothekskatalog der TIB suchen Hier geht es zum klassischen Katalog ↗

Seite 1 von 1.265.680 Ergebnissen Sortieren nach: Relevanz | Aktualität neu zuerst | Titel A-Z

Treffer filtern

Erscheinungsjahr

1800 1850 1900 1950 2000

-

Medientyp

- Buch (717.682)
- Hochschulschrift (265.239)
- Report (138.658)
- Konferenzband (82.820)
- Zeitschrift (56.333)

[+ Mehr anzeigen](#)

TIBKAT x Deutsch x [Auswahl aufheben](#)

« < 1 2 3 4 5 6 7 > »

★ **Ausbildungsverhältnisse in der Textilindustrie**
TIBKAT | Nachgewiesen 1986(1987) - 1992(1993); damit Ersch. eingest. cited: 0 | score: 0.8213173

★ **Magazin der neuesten Erfindungen, Entdeckungen und Verbesserungen der Engländer, Franzosen, Italiener, Amerikaner und deutschen in der gesamten Gewerbekunde**
TIBKAT | N.F. [1.]1818/21 - 4.1829/32 = Nr. 1-36; Neueste F. 1.1832/33 - 5.1839/42[?] cited: 0 | score: 0.8213173

★ **Vorträge, Berichte und Protokolle der Ansprechpartnertagung SEP**



https://github.com/runnwerth/annif_automated_indexing_at_tib/wiki/Definition-des-Daten-Dumps



feedback

SOLR-Anfrage

```
language%5D%5B1%5D=de
supplierPrefix%5D%5B0%5D=tibkat
classification%2F%40classificationProcedure%3Dmapping
```

OAI-Schnittstelle

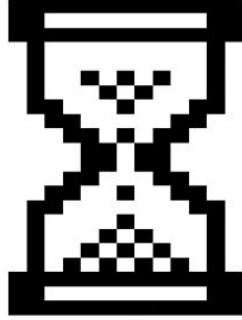
```
set=collection~tibkat
language:de
xmlPath:classification/@classificationProcedure=mapping
```



https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Daten-Dump-herunterladen

3

Datendump
herunterladen





Daten
transformieren



Der fertige Input für annif sieht dann pro Titeldatensatz so aus:

TXT: (Semantisch verwertbare) Metadaten

```
Ende der Hausfrauenehe : Plädoyer gegen e.  
trügerische Existenzgrundlage
```

```
Wiegmann, Barbelies
```

```
Rowohlt
```

```
rororo aktuell : Frauen aktuell
```

```
Hausfrau
```

```
Soziale Situation
```

```
Geschiedene Frau
```

```
Ehe
```

TSV: Ausgewählte Erschließungsdaten

```
<http://unicorn.fly/linsearch#sow> Sozialwissenschaften
```



5

Goldstandard
definieren

6

Testmenge
extrahieren



 https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Anlegen-eines-Goldstandards



7

annif
konfigurieren



- TF-IDF
- FastText in mehreren Varianten
- Omikuji Parabel in mehreren Varianten
- Omikuji Bonsai in mehreren Varianten
- Maui
- vw-multi in mehreren Varianten

- going on...

https://github.com/runnwerth/annif_automated_indexing_at_tib/wiki/Aufsetzen-eines-annif-Projektes



Korrektur zum Vortrag beim annif-Austausch:



1 Minute bis 5 Stunden*



Ergebnisse
analysieren



	Schlechteste Werte	Beste Werte
	TF-IDF	vw-multi (etc)
Precision	0.12	0.66
Recall	0.91	0.88
F1 Score	0.21	0.70



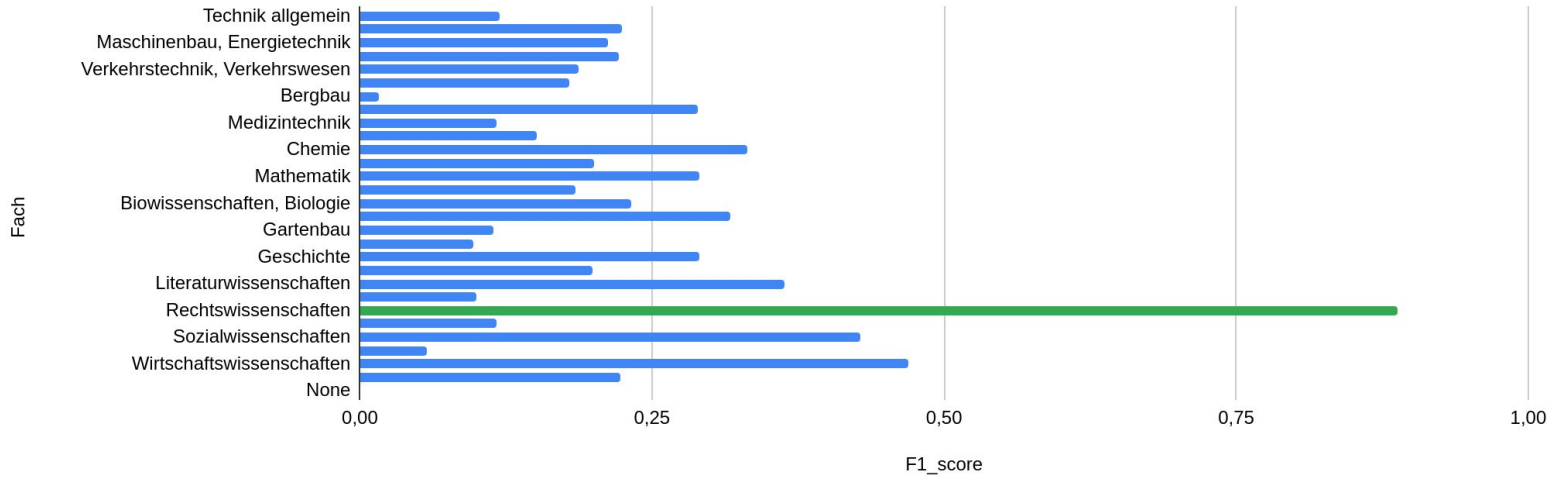


Ergebnisse
analysieren



TF-IDF

Fachabhängiger F1-Score



https://github.com/runnwerth/annif_automated_indexing_at_tib/wiki/Ergebnisse



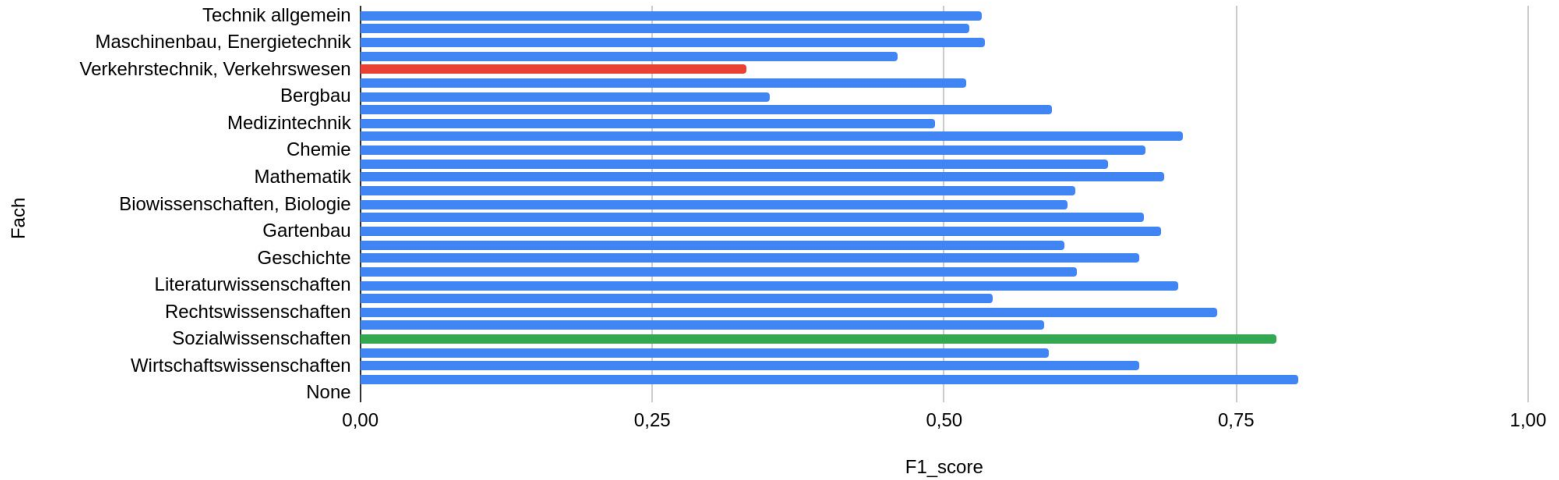


Ergebnisse
analysieren



vw-multi (etc)

Fachabhängiger F1-Score



https://github.com/runwerth/annif_automated_indexing_at_tib/wiki/Ergebnisse



Machen Sie mit!



https://github.com/runnwerth/annif_automated_indexing_at_tib/wiki/

