LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

# Fächerklassifikation mit annif für die Fachfacetten des TIB-Portal

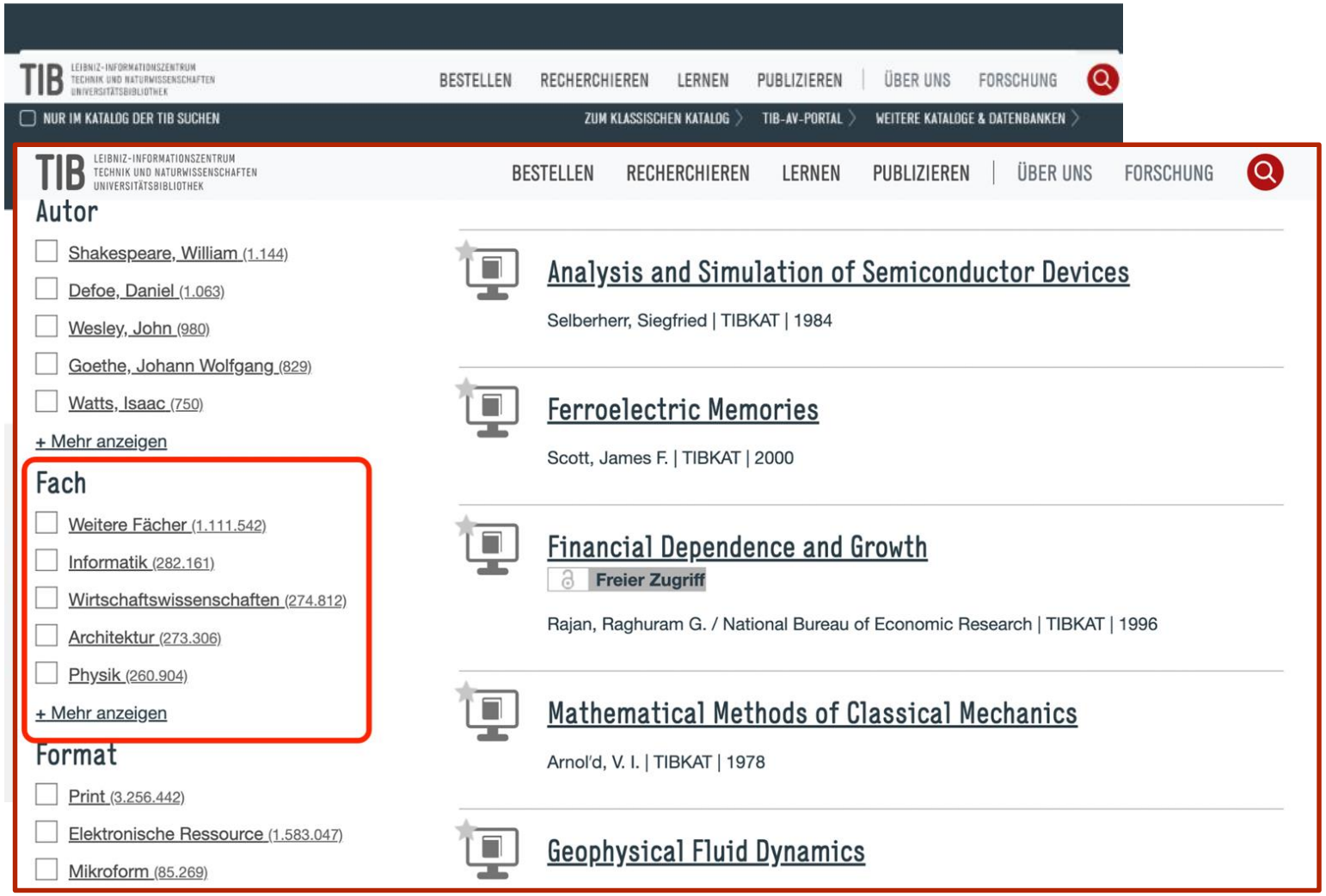
Dr. Holger Israel
Frankfurt/Main, 2022-11-03
Workshop zu KI in Bibliotheken, DNB

Leibniz Gemeinschaft

# Agenda

1. Kontext: Die Fachfacetten im TIB-Portal
   - LinSearch: Status quo
2. Ablösung von Averbis durch **Annif**
   - Warum Annif?
3. Stand des Projekts
   - Trainingsdaten
   - Test- und Validierungsplan
   - Kriterien und Metriken
4. Ausblick

# Fachzuordnung im TIB-Portal



# LinSearch:
# Algorithmus zur Fächerzuordnung

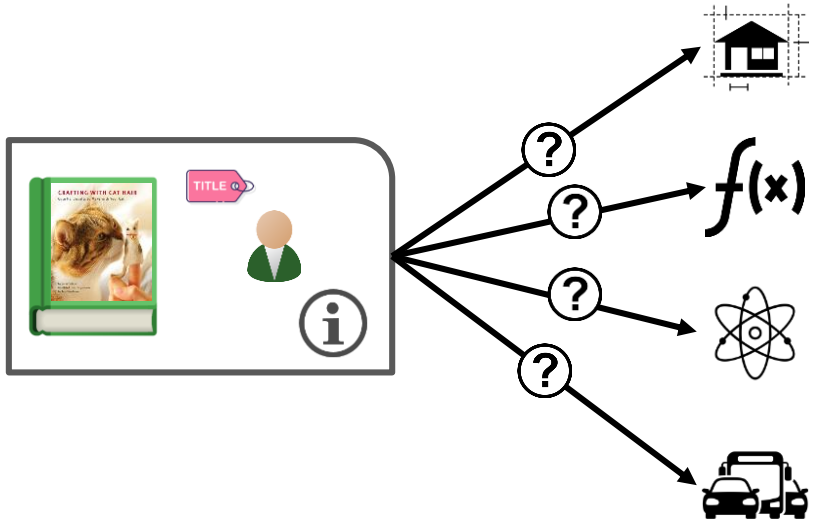
Automatische Fachzuordnung von Veröffentlichungsnachweisen anhand ihrer Metadaten […] in einem vierstufigen Verfahren.

**Stufe 0: Mapping** Zuordnung nach vorhandenen Klassifikationselementen gemäß einer Konkordanz

**Stufe 1: Lexikon** Zuordnung nach fachlich einschlägigen Zeitschriften und Reihen nach ISBN / ISSN

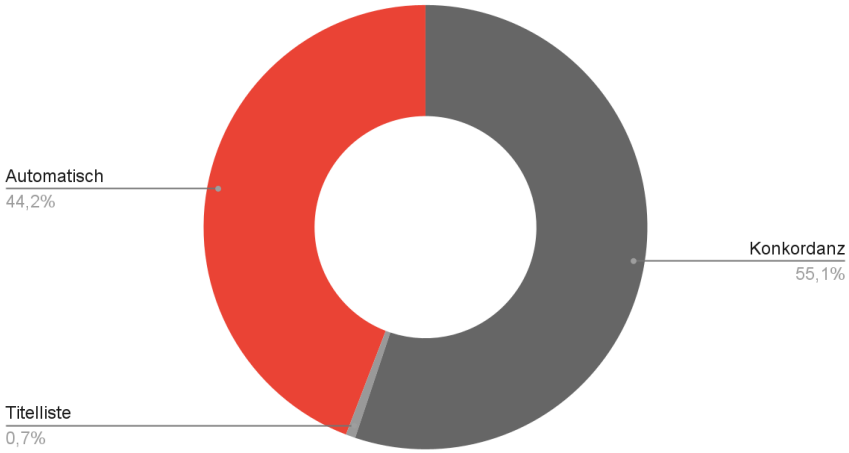
**Stufe 2: Datenlieferant** Zuordnung nach Zugehörigkeit zu einer fachlich einschlägigen Datenkollektion

**Stufe 3: Automatisch** Automatische Fachzuordnung nach linguistischer Indexierung nach Averbis

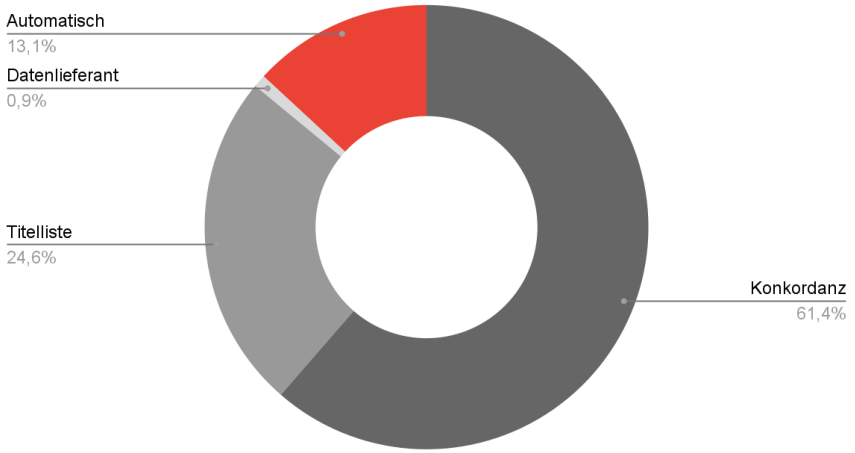


# LinSearch:
# Algorithmus zur Fächerzuordnung

TIBKAT (5 Mio. Datensätze)




TIB-Index (138 Mio. Datensätze)




230 Datensätze nicht zugeordnet:
Bestände von 1955 - 1970

381 Datensätze nicht zugeordnet:
151 EPA,
230 TIBKAT (1955 - 1970)

# Ablösung von Averbis durch Annif: Nachteile Averbis

- Kosten: jährlich fünfstelliger Euro-Betrag für die bloße Nutzung
- Closed source: veraltetes Verfahren basierend auf Support Vector Machine (mittlerweile gibt es dafür angepasste neuronale Netze)
  - Konfiguration ist eine Black Box
  - Ergebnisse nicht nachvollziehbar
- Closed source: Kein Einfluß auf die Weiterentwicklung
- Closed source: Keine Anpassungen (Training, Updates, etc.) verfügbar
- Closed source: Keine Interaktionsfunktionalität bei Fehlern oder schlechten Ergebnissen
- Umfang: nur TIB-Kernfächer erfasst (seit 2016 Untergliederung der Technikfächer)

# Ablösung von Averbis durch Annif: Vorteile Annif

- Freie Software: Keine Kosten für Lizenzgebühren
- Open Source: Enger Community-Austausch mit bereits etablierten und im Wachsen begriffenen Austauschstrukturen. **Zum Beispiel auf diesem Workshop!**
  - Modell- und Datenaustausch ist jedoch nicht ohne Weiteres möglich oder sinnvoll ➔ Sinnvolle Aufgabe für diese Runde
- Open Source: Freie Anpassungs- und Konfigurationsmöglichkeiten. Neue Verfahren können einfach integriert werden.
- Open Source: TIB wird ihrer Vorbildrolle in der Open Science – Transformation gerecht
- Umfang: Weitere Anwendungsmöglichkeiten (bzw. bereits konkrete Anfragen), insb. für verbale Sacherschließung (z.B. Vorschlagstool)

# Bisherige Experimente

- Baut auf dem für Averbis etablíerten Workflow auf. ➔ **Vergleichbarkeit**
- Testdatensätze müssen nicht zwingend einen **Abstract** enthalten
- Grundlage: Metadaten aus TIBKAT, für die auch BKs vorhanden sind
- Vergleich mit bekannter Averbis-Klassifikation (relative Qualität)
- Ohne Aussage / Kontrolle der Qualität der Averbis-Klassifikation!

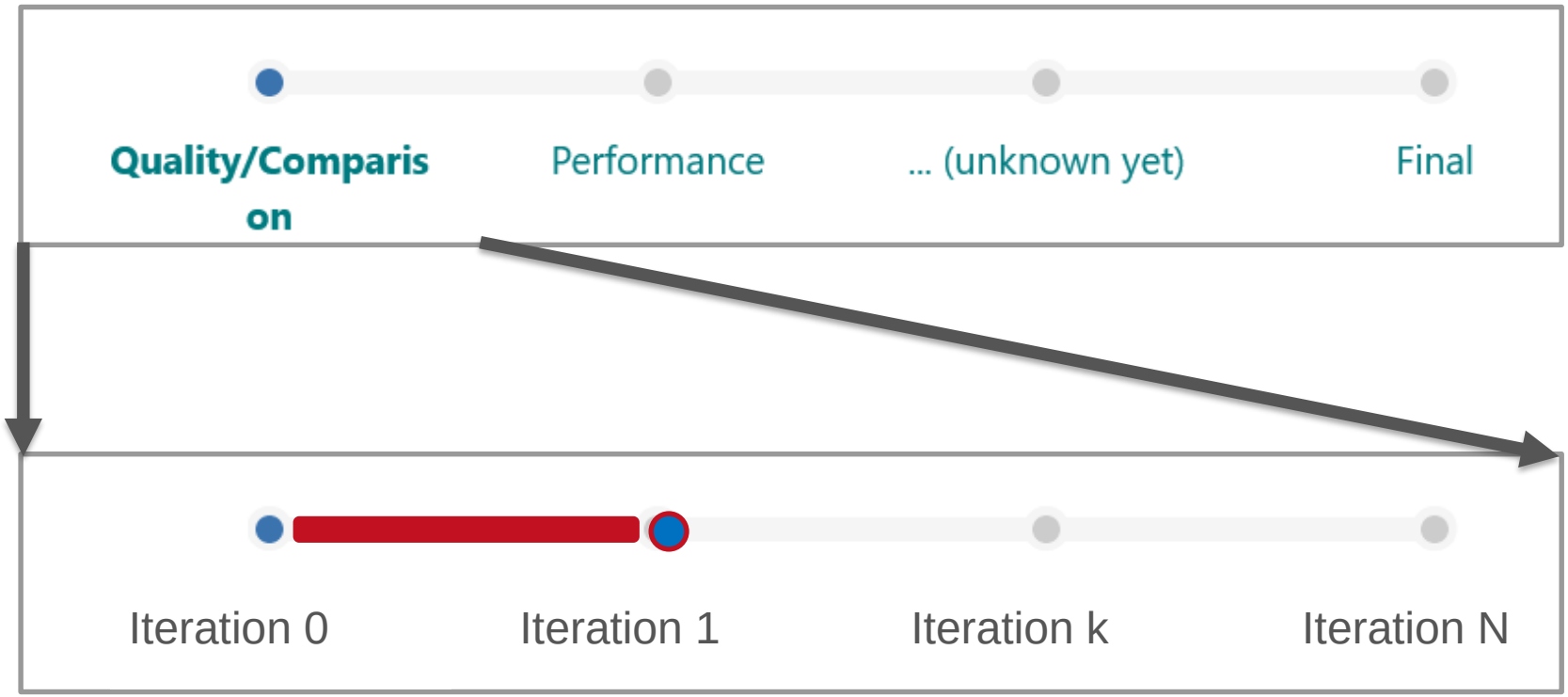
- **Algorithmen / “Projekte”**
  - Tib-linsearch-fasttext (en/de)
  - Omikuji-parabel (en)
  - Linsearch-svc (en)

# Bisherige Ergebnisse

Beispiel **omikuji-parabel-en**, *threshold* = 0.5:

- **Exact match: 38.7%:** Averbis und annif geben übereinstimmende Klassifizierungen zurück.
- **Subset match: 11.0%:** Annif gibt eine Teilmenge der Averbis-Klassifizierungen zurück.
- **Superset match 1.6%**: Annif gibt neben den Averbis-Klassifizierungen zusätzliche Ergebnisse zurück.
- **Partial match: 0.3 %:** Die Ergebnismengen von Annif und Averbis überschneiden sich nur teilweise.
- **“Wrong match”: 3.8 %:** Die Ergebnismengen von Annif und Averbis überschneiden sich gar nicht.
- **None: 44.5 %:** Keine Klassifikation durch Annif, aber durch Averbis

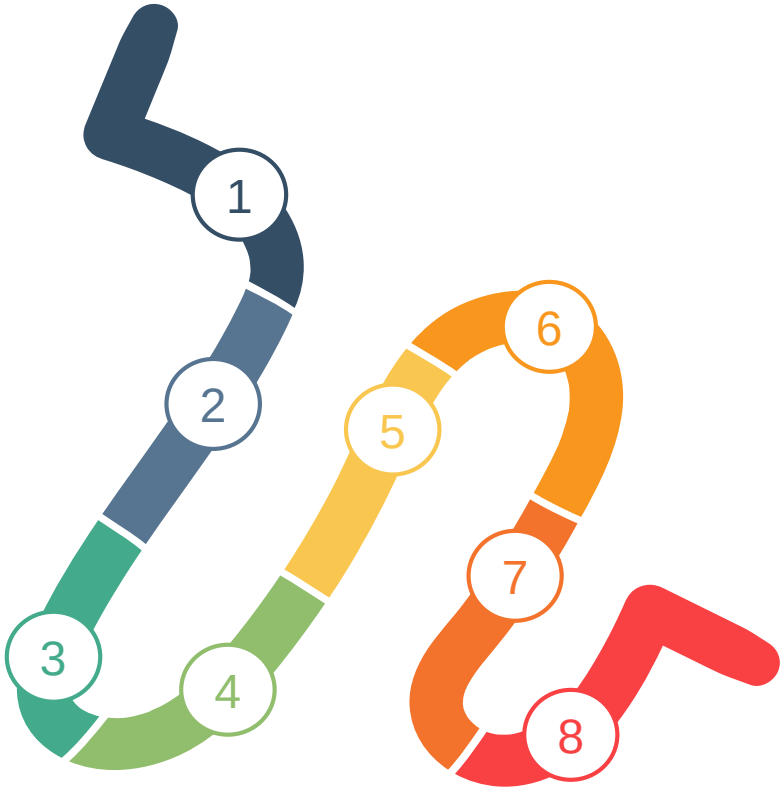
# Stand des Projekts




Iteratives Vorgehen: Verbesserung der Testdaten, Algorithmen, Hyperparameter

# Schema einer Iteration

Jede Iteration setzt sich aus den folgenden Schritten zusammen:




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

# Stand des Projekts:
# Kriterien und Metriken

Definition eines Anforderungsprofils im wissenschaftlichen Dienst

- A: Relative Qualität (d.h. Vergleich zu Averbis)
  - Ziele mit Zahlenwerten definiert
- B: Optimierbarkeit
- C: Absolute Qualität (Vergleich gegen eine Stichprobe mit intellektueller Sacherschließung)
  - Ziele (F-Score) mit Zahlenwerten definiert
- D: Formalien: So genau wie nötig, so flexibel wie möglich…

- Hinzu kommen technische Anforderungen (Formate, Geschwindigkeit, …)

# Stand des Projekts:
# Kriterien und Metriken

## A: Relative Qualität (d.h. Vergleich zu Averbis)

Wir definieren als korrekt klassifiziert relativ zu Averbis solche Werke, die von Annif denselben Fächern zugordnet werden wie durch Averbis (*exact matches*) oder zu einer Auswahl derselben Fächer wie durch Averbis (*subset matches*).

- **Kriterium A1**: Über alle Fächer gemittelt muß die Rate der korrekt relativ zu Averbis klassifizierten Werke bei mindestens 80 % liegen.
- **Kriterium A2**: In jedem Fach muß die Rate der korrekt relativ zu Averbis klassifizierten Werke bei mindestens 70 % liegen.
    - *Anmerkung: Dies könnte eine recht einschränkende Bedingung sein. Manche Fächer könnten an sich schwieriger festzustellen sein; zudem haben die Fächer eine deutlich verschiedene a priori – Häufigkeit. Evtl. sollten wir von vornherein definieren, für welche Fächer dies nicht gelten soll.*
- **Kriterium A3**: Über alle Fächer gemittelt darf die Rate der relativ zu Averbis falsch klassifizierten Werke (*wrong matches*) bei höchtens 10 % liegen.
- **Kriterium A4**: Über alle Fächer gemittelt darf die Rate der von Averbis aber nicht von annif klassifizierten Werke (none) bei höchstens 5 % liegen.

# Auswahl der Trainingsdaten

**Ursprüngliche Kriterien**

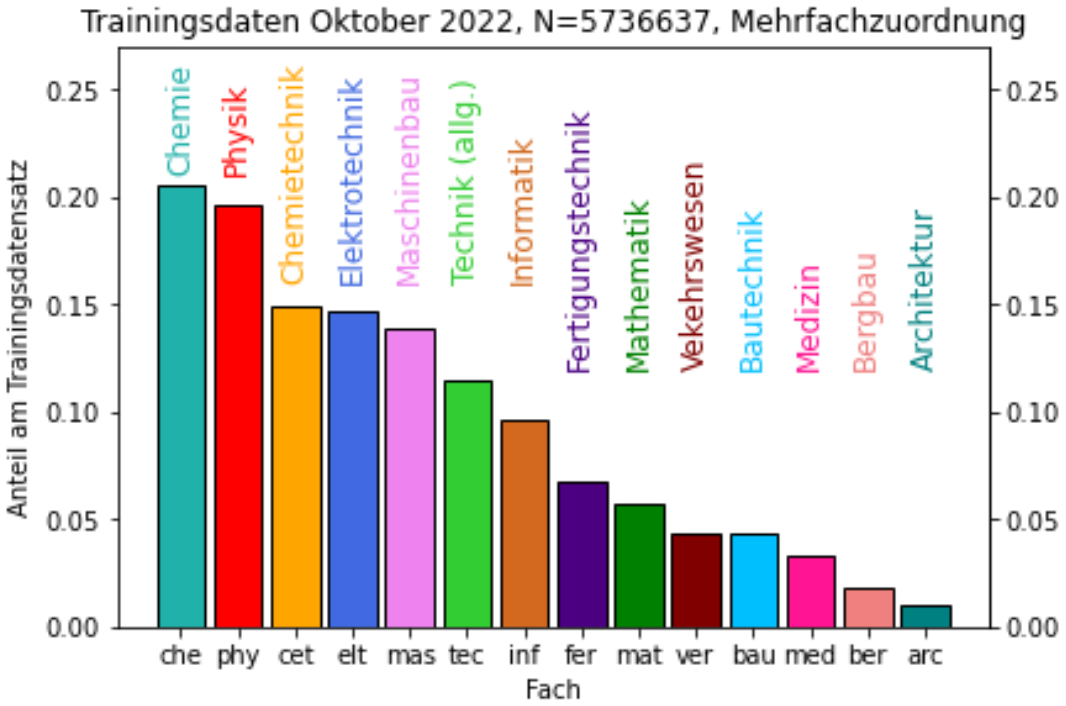
- Grundgesamtheit: TIBKAT-Kollektion
- Nur Dokumente mit gültiger LinSearch-Klassifikation
- Klassifiziert in der Mapping-Stufe (Stufe 0)
- Aus historischen Gründen: *mat* und *inf* bei Publikationsjahr <1995 ignorieren

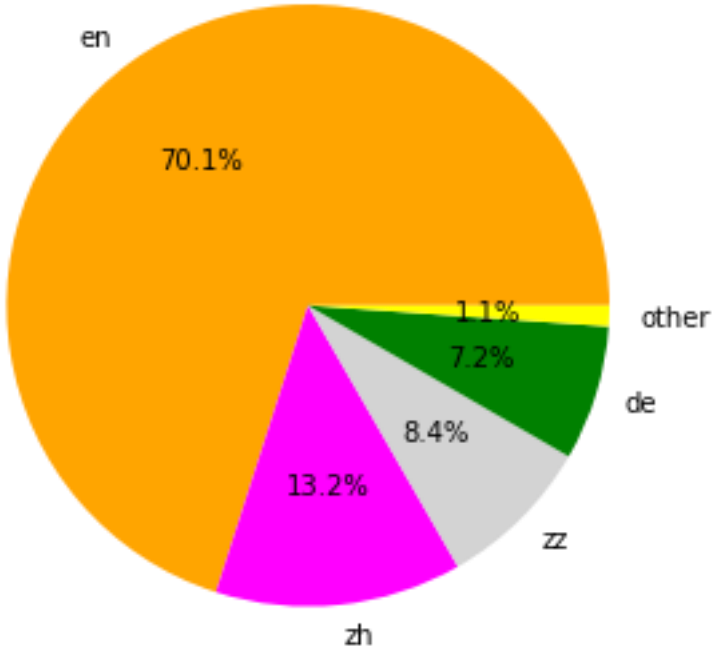
**Aktuelle Kriterien**

- Grundgesamtheit: Kompletter TIB-Index
- Nur Dokumente mit gültiger LinSearch-Klassifikation
- Klassifiziert in der Mapping-Stufe (Stufe 0)
- Fachübergreifend: Publikationsjahr ≥1995 (repräsentativ für Neuzuwachs)

# Trainingsdatensatz Oktober 2022

T…




Bei weitem häufigste Datenquelle („publisher“): Europäisches Patentamt, 46%

- Vermutlich wenig nützlich für das Training
- Erklärt auch den hohen Anteil chinesischsprachiger Quellen

# Sprachfragen

- Aus TIB-Sicht sind vorranging englisch- und deutschsprachige Dokumente für das TIB-Portal relevant.
- Auch prinzipiell mehrsprachige Annif-Projekte sind für eine Sprache optimiert
- ➔ Relativ unabhängige Suche nach je einem Algorithmus für Englisch und Deutsch
- Im Prinzip können sich die Sprachen von Metadaten und Volltext einer Quelle unterscheiden.
- Oft fehlt der Sprachcode**: Kann Annif selbst hier helfen?**

TIBKAT:01001487X: phy

</doc><doc boost="**1.0**"><field name="**document_title**">Quanten-Hall-Effekt</field><field name="**document_name**">TIBKAT:01001487X</field><field name="**corporate_creator**">Physikalisch-Technische Bundesanstalt</field><field name="**language**">zz</field></doc>

# Ausblick

**Nächste Experimente** nach Finalisierung des aktuellen Testdatensatzes:

- **Experiment 1**: **Extremfall „*threshold* =0“:** Jedes Werk bekommt das Fach mit dem höchsten Score zugewiesen. (Keine Mehrfachzuordnungen, d.h. *limit=1*)
  - Wie hoch ist der Anteil der im Vergleich zu Averbis falsch zugewiesenen Werke?
- **Experiment 2: Feststellen der Basislinie:** Nun lassen wir Mehrfachzuordnungen vor und setzen *limit=5.*
  - Max. 4 Fächer pro Werk mit Averbis
  - Informationsgehalt pro Zuordnung sinkt mit der Anzahl
  - ➔Wie gehen wir im folgenden mit Mehrfachzuordnungen um?

**Experiment 3: Training über Metadaten mit vs. ohne Abstract**

**Experiment 4: Hauptversuch: Wieviel fehlt uns noch, um die definitieren Kriterien zu erreichen?**

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

# Beteiligte Koautor:innen:


Susanne Arndt

Christine Baumgarten

Jacopo De Benedetto

Berrit Genat

Holger Israel

Mila Runnwerth (ehemals)




Leibniz Gemeinschaft

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

# MEHR INFORMATIONEN

www.tib.eu

**Kontaktdaten**
Dr. Holger Israel
T 0511 762-3979, holger.israel@tib.eu



Leibniz Gemeinschaft