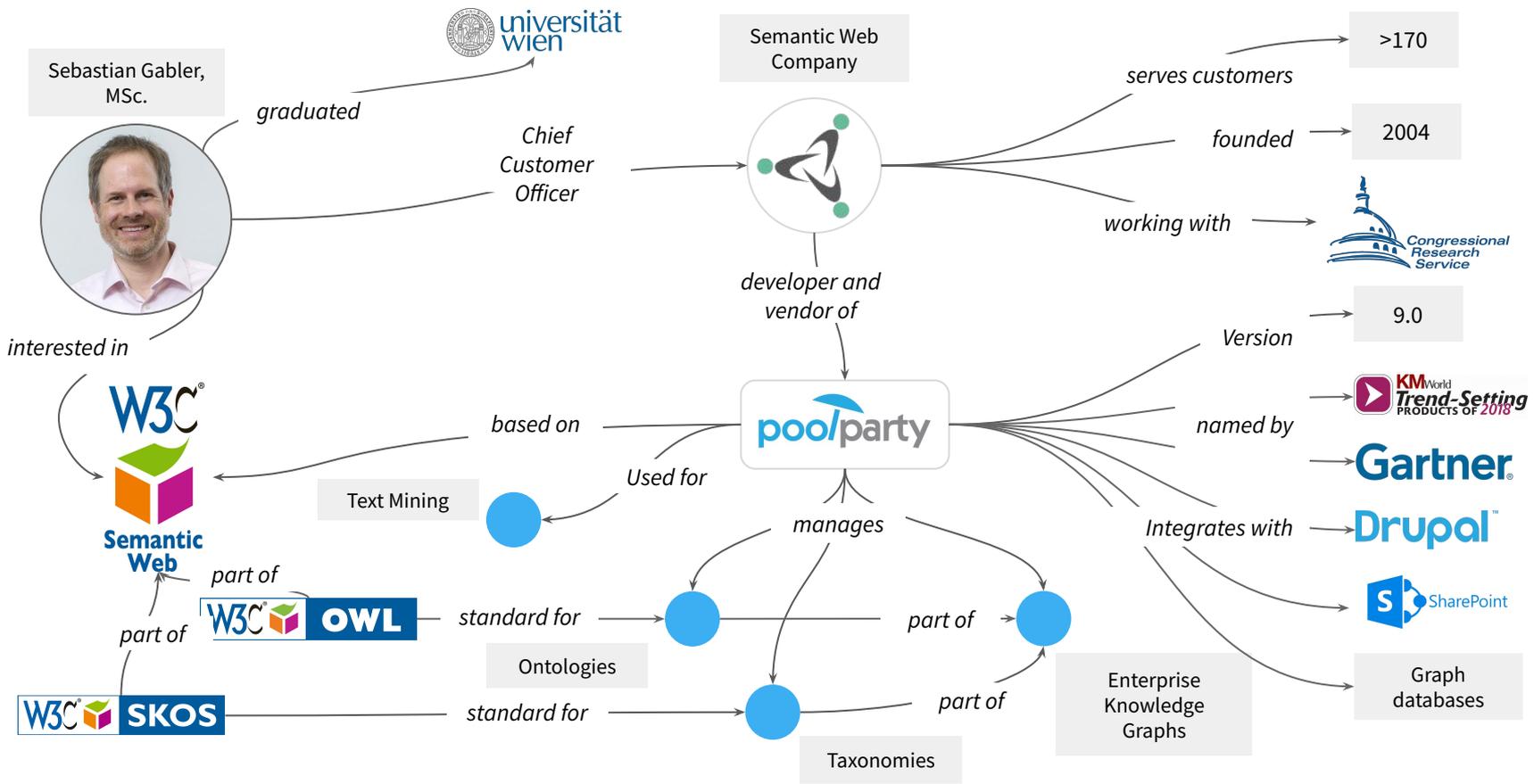


# Vergabe von DDC-Sachgruppen mittels eines Schlagwort-Thesaurus

Dipl. Tonm. Sebastian Gabler, MSc.  
[sebastian.gabler@gmx.net](mailto:sebastian.gabler@gmx.net)



# Agenda

- Warum braucht Information Retrieval konsistente Metadaten?
- Automatisierung unter Nutzung einer Beschreibungslogik
- Klassen- und dokumentenzentrische Herausforderungen
- Evaluierung des Verfahrens
- Zusammenfassung und Ausblick
- Diskussion

*“Information Retrieval  
benötigt konsistente  
und präzise Metadaten.”*

# Stichwortsuche führt zu vielen Resultaten

→ Über die Deutsche  
Nationalbibliothek

## Ergebnis der Suche nach: *"flughafen"* im Bestand: Gesamter Bestand

1 - 10 von 3436

Datum (neuestes zuerst) ▼

sortieren →



- 1 Flughafen Wimmelbuch  
Walther, Max. - Berlin : adrian & wimmelbuchverlag, 2022, 1. Auflage

# Schlagwortsuche erlaubt hohe Präzision

→ Über die Deutsche  
Nationalbibliothek

**Ergebnis der Suche nach: *swiRef=041547527 sortBy jhr/sort.descending***  
**im Bestand: Gesamter Bestand**

1 - 10 von 408

Datum (neuestes zuerst) ▼

sortieren →



- 1 [Alles über Flugzeuge]  
Repülővel utazunk  
Metzger, Wolfgang. - [Budapest] : Scholar Kid, [2021], Második kiadás

# Facettierte Suche nach DDC-Sachgruppen

- Unterteilung des Publikationsaufkommens nach DDC (Alle Reihen der Deutschen Nationalbibliografie)
- 104 Sachgruppen
- DDC als alternative Systematik für GND nutzbar
- Notationen nach den obersten 3-5 DDC-Klassen
- Gute Ergänzung für DDC- Kurznotationen
- DDC ist als SKOS-Taxonomie publiziert
- DDC bietet ähnliche Zugänge in weiteren Suchmaschinen anderer Communities und Sprachräume (VIAF, BASE, EBSCO, Worldcat, ...)



The screenshot shows a web browser window with the URL <https://portal.dnb.de/opac/checkCategory?categoryId=sg320>. The page displays a list of DDC categories with checkboxes for selection. The categories are organized into sections: **Philosophie und Psychologie**, **Religion**, **Sozialwissenschaften**, and **Sprache**. The category **320 Politik** is selected, indicated by a checked checkbox. Other categories include 020 Bibliotheks- und Informationswissenschaft, 030 Enzyklopädien, 050 Zeitschriften, fortlaufende Sammelwerke, 060 Organisationen, Museums- und Bibliothekswissenschaft, 070 Nachrichtenmedien, Journalismus, Verlagswesen, 080 Allgemeine Sammelwerke, 090 Handschriften, seltene Bücher, 100 Philosophie, 130 Parapsychologie, Okkultismus, 150 Psychologie, 200 Religion, Religionsphilosophie, 220 Bibel, 230 Theologie, Christentum, 290 Andere Religionen, 300 Sozialwissenschaften, Soziologie, Anthropologie, 310 Allgemeine Statistiken, 320 Politik, 330 Wirtschaft, 333.7 Natürliche Ressourcen, Energie und Umwelt, 340 Recht, 350 Öffentliche Verwaltung, 355 Militär, 360 Soziale Probleme, Sozialdienste, Versicherungen, 370 Erziehung, Schul- und Bildungswesen, 380 Handel, Kommunikation, Verkehr, 390 Bräuche, Etikette, Folklore, 400 Sprachen, Linguistik, 420 Englisch, and 430 Deutsch.

# Metasuche und Facettierung über Klassifikation

*WebDewey Search* Suche mit der Dewey-Dezimalklassifikation

Suchbegriff oder Notation:    Kürzungsstriche (DDC-Kurznotationen) anzeigen

Suche in:  DNB  GBV  HeBIS  SUB  SWB  FUB

## Haupttafeln

Notation	Thema	Titel in dieser Klasse	Titel in dieser Klasse und Unterklassen
	<a href="#">Haupttafeln</a>		
300	<a href="#">Sozialwissenschaften</a>	0 (DNB) <a href="#">28 (GBV)</a> 0 (HeBIS) <a href="#">12 (SUB)</a> <a href="#">204 (SWB)</a> <a href="#">6098 (FUB)</a>	<a href="#">438935 (DNB)</a> <a href="#">2174504 (GBV)</a> <a href="#">2537399 (HeBIS)</a> <a href="#">508441 (SUB)</a> <a href="#">1418551 (SWB)</a> 0 (FUB)
380	<a href="#">Handel, Kommunikation &amp; Verkehr</a>	0 (DNB) <a href="#">3 (GBV)</a> 0 (HeBIS) <a href="#">1 (SUB)</a> <a href="#">3 (SWB)</a> <a href="#">187 (FUB)</a>	<a href="#">20603 (DNB)</a> <a href="#">133423 (GBV)</a> <a href="#">156817 (HeBIS)</a> <a href="#">17127 (SUB)</a> <a href="#">38608 (SWB)</a> <a href="#">14014 (FUB)</a>
383-388	<a href="#">Kommunikation und Verkehr</a>	<a href="#">316 (DNB)</a> <a href="#">2097 (GBV)</a> <a href="#">97090 (HeBIS)</a> <a href="#">249 (SUB)</a> <a href="#">2688 (SWB)</a> k.A. (FUB)	<a href="#">13321 (DNB)</a> <a href="#">24476 (GBV)</a> <a href="#">97090 (HeBIS)</a> <a href="#">3751 (SUB)</a> <a href="#">19197 (SWB)</a> k.A. (FUB)
387	<a href="#">Schifffahrt, Luft-, Weltraumverkehr</a>	<a href="#">12 (DNB)</a> <a href="#">176 (GBV)</a> 0 (HeBIS) <a href="#">27 (SUB)</a> <a href="#">140 (SWB)</a> <a href="#">54 (FUB)</a>	<a href="#">2784 (DNB)</a> <a href="#">5937 (GBV)</a> <a href="#">9989 (HeBIS)</a> <a href="#">668 (SUB)</a> <a href="#">3280 (SWB)</a> <a href="#">812 (FUB)</a>
387.7	<a href="#">Luftverkehr</a>	<a href="#">60 (DNB)</a> <a href="#">221 (GBV)</a> 0 (HeBIS) <a href="#">24 (SUB)</a> <a href="#">260 (SWB)</a> <a href="#">63 (FUB)</a>	<a href="#">1027 (DNB)</a> <a href="#">1851 (GBV)</a> <a href="#">4194 (HeBIS)</a> <a href="#">214 (SUB)</a> <a href="#">1450 (SWB)</a> <a href="#">351 (FUB)</a>
387.73	<a href="#">Luftfahrzeuge und Einrichtungen</a>	<a href="#">7 (DNB)</a> <a href="#">11 (GBV)</a> 0 (HeBIS) <a href="#">1 (SUB)</a> <a href="#">29 (SWB)</a> <a href="#">2 (FUB)</a>	<a href="#">470 (DNB)</a> <a href="#">471 (GBV)</a> <a href="#">1703 (HeBIS)</a> <a href="#">50 (SUB)</a> <a href="#">363 (SWB)</a> <a href="#">44 (FUB)</a>
<b>387.736</b>	<b>Flughäfen</b>	<a href="#">41 (DNB)</a> <a href="#">106 (GBV)</a> 0 (HeBIS) <a href="#">10 (SUB)</a> <a href="#">133 (SWB)</a> <a href="#">18 (FUB)</a>	<a href="#">240 (DNB)</a> <a href="#">377 (GBV)</a> <a href="#">589 (HeBIS)</a> <a href="#">46 (SUB)</a> <a href="#">276 (SWB)</a> <a href="#">42 (FUB)</a>
<a href="#">387.7362</a>	<a href="#">Einrichtungen</a>	<a href="#">16 (DNB)</a> <a href="#">19 (GBV)</a> 0 (HeBIS) 0 (SUB) <a href="#">11 (SWB)</a> 0 (FUB)	<a href="#">18 (DNB)</a> <a href="#">23 (GBV)</a> <a href="#">71 (HeBIS)</a> 0 (SUB) <a href="#">14 (SWB)</a> 0 (FUB)
<a href="#">387.7364</a>	<a href="#">Teilbereiche und Dienste</a>	<a href="#">17 (DNB)</a> <a href="#">28 (GBV)</a> 0 (HeBIS) <a href="#">2 (SUB)</a> <a href="#">24 (SWB)</a> <a href="#">4 (FUB)</a>	<a href="#">40 (DNB)</a> <a href="#">43 (GBV)</a> <a href="#">35 (HeBIS)</a> <a href="#">5 (SUB)</a> <a href="#">30 (SWB)</a> <a href="#">6 (FUB)</a>

DNB = Deutsche Nationalbibliothek | GBV = Gemeinsamer Bibliotheksverbund | HeBIS = HeBIS Verbundkatalog | SUB = SUB Göttingen | SWB = Südwestdeutscher Bibliotheksverbund | FUB = FU Berlin

*“Sacherschließung benötigt agile  
Prozesse.”*

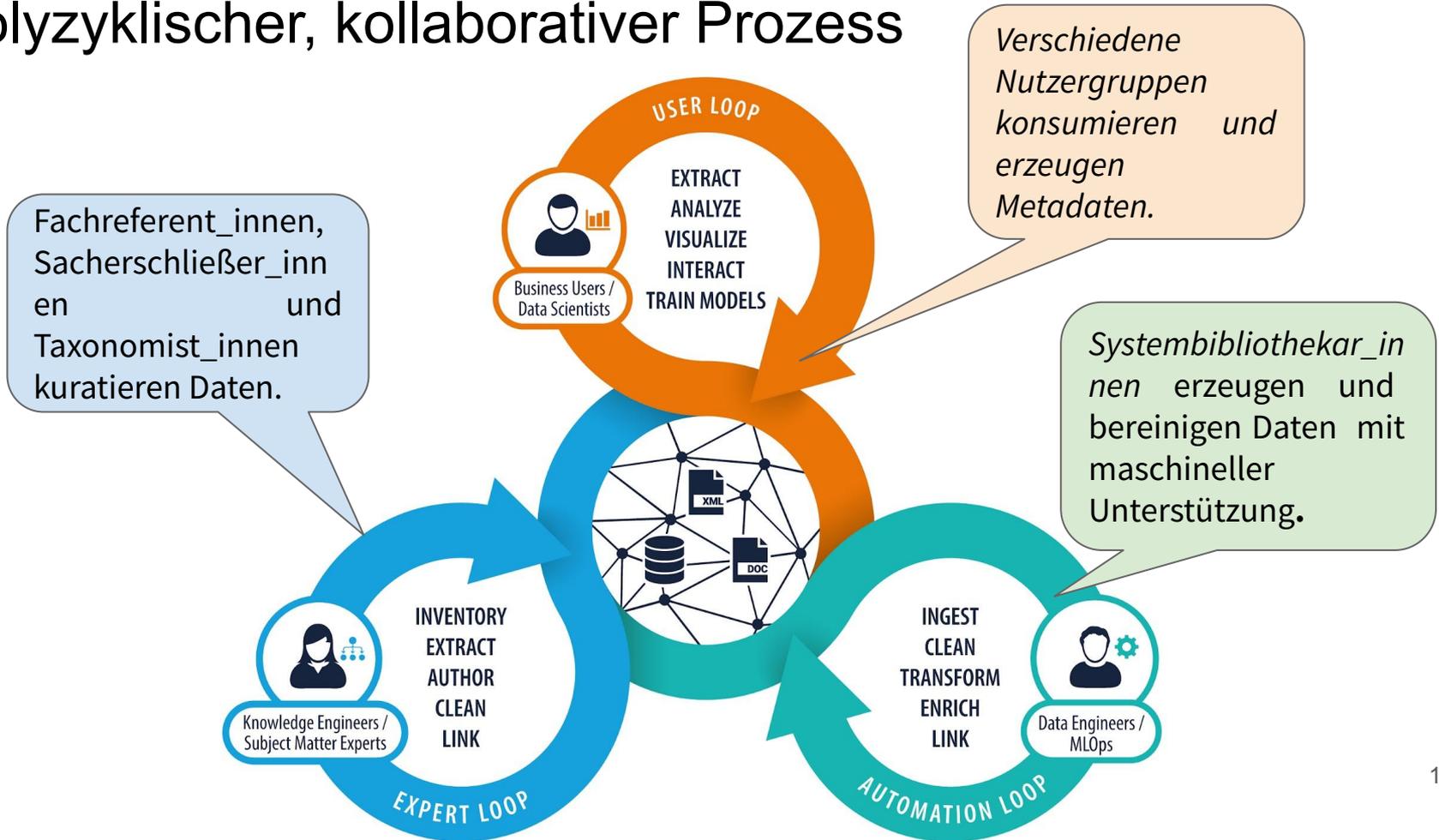
# Aktuelle Anforderungen an Sacherschließung

- Starker Zuwachs im Bestand, vor allem durch digitale Ressourcen\*
- Metadaten unterstützen aktuelle Retrievalkonzepte\*
- Transparente Prozesse, einschließlich der maschinell getroffenen Entscheidungen\*
- Nachnutzung zugelieferter Metadaten\*\*
- Sacherschließung als agiler Prozess (“zyklische Erschließung”)\*\*
- Harmonisierung und Verbesserung der zugelieferten Metadaten\*\*
- Intellektuelle Kontrolle der maschinell erzeugten Metadaten\*\*
- Daten sollen *FAIR* (findable, accessible, interoperable, reusable) sein
- Responsible AI Direktive der EU (2022 wirksam)

\* PETRUS - Prinzipien, 2009

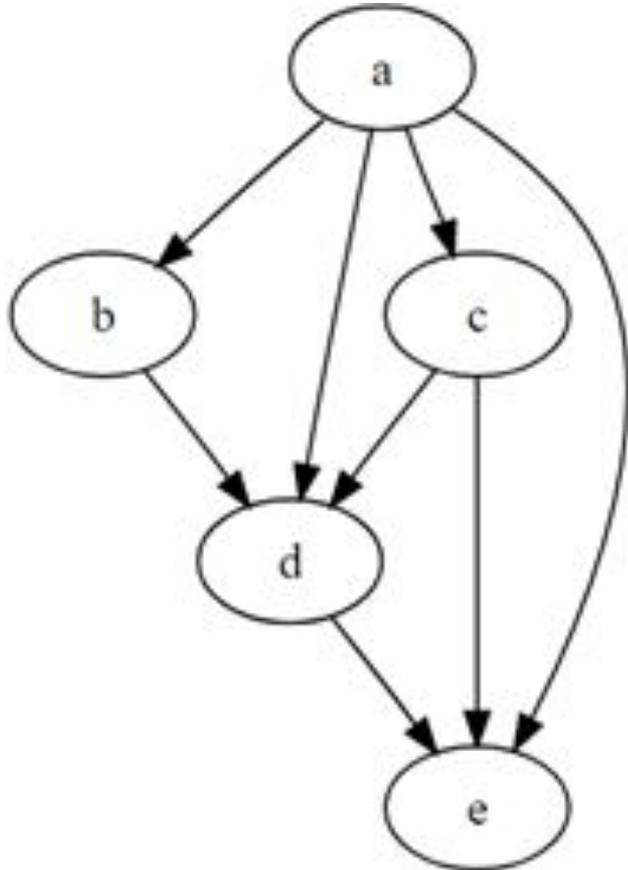
\*\* Positionspapier der DNB, 2017

# Polyzyklischer, kollaborativer Prozess



*“Das Verfahren ermittelt die gefundenen Schlagwörter über Vergleich der Zeichenfolgen Benennungen im Thesaurus, ordnet die Treffer nach ihrer Relevanz im Text und gibt die zugeordneten Sachgruppen rangordnend zurück.”*

# Transitiver Schluss im azyklischen, gerichteten Graphen



Es gilt:

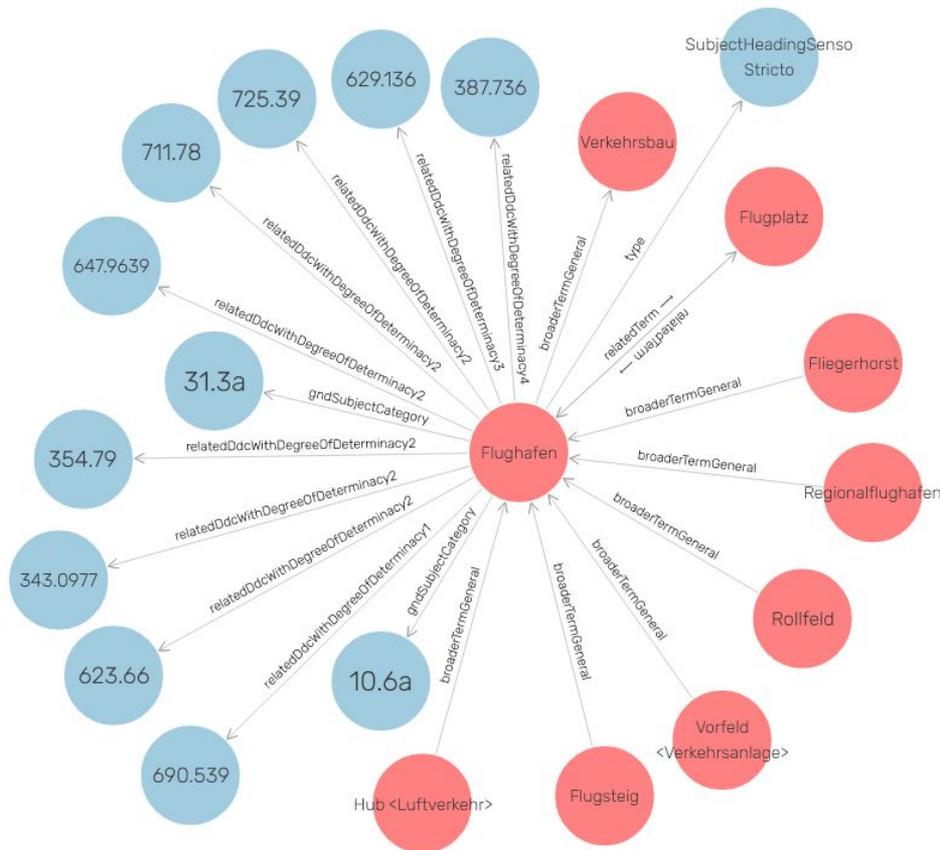
a ist Oberbegriff von b, ist Oberbegriff von c, ist Oberbegriff von e.

Es gilt daher auch:

a ist Oberbegriff von e.

-> Wenn Schlagwörter Unterbegriffe einer Sachgruppe sind, gehört das Dokument in die betreffende Sachgruppe

# Beschreibungslogik GND Sachschlagwörter



## Flughafen

 Flughafen

Types:

<https://d-nb.info/standards/elementset/gnd>

RDF rank:

0

<https://d-nb.info/standards/elementset/gnd#gndIdentifier>  
**4154752-4**

<https://d-nb.info/standards/elementset/gnd#oldAuthorityNumber>  
**(DE-588c)4154752-4**

<https://d-nb.info/standards/elementset/gnd#preferredNameForTheSubjectHeading>  
**Flughafen**

<https://d-nb.info/standards/elementset/gnd#preferredNameForTheSubjectHeading>

# Falsche Polyhierarchie / GND Systematik

The screenshot displays the GND interface. On the left, a hierarchical tree lists various concepts, with 'Fliegerhorst (2)' highlighted in yellow. The main panel shows the details for 'Fliegerhorst' (https://d-nb.info/gnd/4071257-6). The interface includes a top navigation bar with 'PROJECT', 'CORPORA', 'TOOLS', and 'ADVANCED' tabs, and a search bar containing 'de' and 'Flughafen'. The main content area is divided into several sections:

- SKOS**: A section for SKOS concepts, showing a list of related terms.
- Broader Concepts**: A list of concepts that are broader than 'Fliegerhorst', including 'Flughafen' and 'Militär'.
- Narrower Concepts**: A list of concepts that are narrower than 'Fliegerhorst', including 'Fliegerhorst Detmold' and 'Fliegerhorst Mönkeloh'.
- Related Concepts**: A list of concepts related to 'Fliegerhorst'.
- Top Concept of Concept Schemes**: A list of top concepts of concept schemes.
- Exact Matching Concepts**: A list of concepts that exactly match 'Fliegerhorst'.
- Close Matching Concepts**: A list of concepts that closely match 'Fliegerhorst'.

On the right side of the main panel, there are several tabs: 'Details', 'Notes', 'Documents', 'Linked Data', 'Triples', 'Visualization', 'Quality Management', and 'History'. Below these tabs, there are several sections for managing the concept:

- Preferred Label**: 'Fliegerhorst' (de)
- Alternative Labels**: 'Airbase' (de), 'Luftwaffenstützpunkt', 'Militärflughafen', 'Militärflugplatz'.
- Hidden Labels**: A plus sign (+) to add hidden labels.
- Notation**: A plus sign (+) to add notation.
- Scope Notes**: A plus sign (+) to add scope notes.
- Example**: A plus sign (+) to add an example.

# Lösung: DDC Sachgruppen als upper ontology

PROJECT CORPORA TOOLS ADVANCED de Flughafen

Flugbetrieb (0)  
Flugboot (0)  
Flughafen (0)  
Flughafenbetrieb (0)  
Flughafengebäude (0)  
Flugingenieur (0)  
Fluglarm (0)  
Fluglinie (0)  
Flugmarke (0)  
Flugpassagier (0)  
Flugplan (0)  
Flugplanung (0)  
Flugplatz (0)  
Flugpreis (0)  
Flugreise (0)  
Flugroute (0)  
Flugsicherheit (0)  
Flugsicherheitsbegleiter (0)  
Flugsicherung (0)  
Flugsteig (0)  
Flugticket (0)  
Flugzeug (0)  
Flugzeugführer (0)  
Flugzeugführerin (0)  
Flugzeughalle (0)  
Flugzeugmarkt (0)  
Flüssiggashandel (0)  
Flussmeister (0)

## Flughafen

[https://d-nb.info/gnd/4194792-4](#)

+ Add to Collection ⌚ Add to Blacklist ⌚ Add to ExactMatch 🗑 Delete Concept

Details Notes Documents Linked Data Triples Visualization Quality Management History

SKOS 👤 +

**Broader Concepts**

- ⊗ [340 - Recht](#)
- ⊗ [350 - Öffentliche Verwaltung](#)
- ⊗ [380 - Handel, Kommunikation, Verkehr](#)
- ⊗ [620 - Ingenieurwissenschaften und Maschinenbau](#)
- ⊗ [640 - Hauswirtschaft und Familienleben](#)
- ⊗ [710 - Landschaftsgestaltung, Raumplanung](#)
- ⊗ [720 - Architektur](#)

**Narrower Concepts**

🔍 +

**Related Concepts**

🔍

**Top Concept of Concept Schemes**

🔍

**Exact Matching Concepts**

- ⊗ [http://zbw.eu/stw/descriptor/13516-0](#)

**Preferred Label**

⊗ Flughafen de

**Alternative Labels**

- ⊗ Airport de
- ⊗ Verkehrsflughafen

+

**Hidden Labels**

+

**Notation**

+

**Scope Notes**

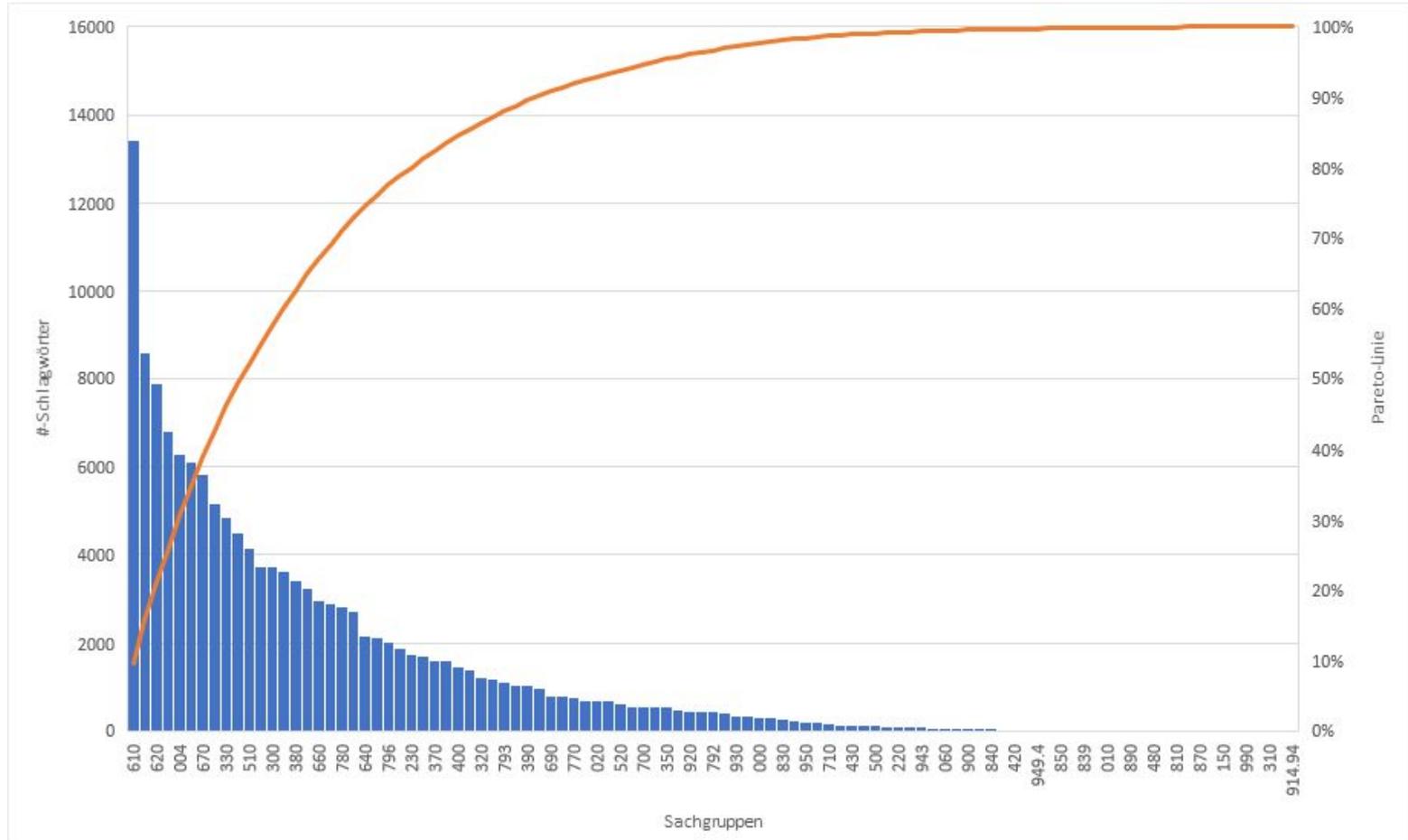
+

**Example**

+

**Definitions**

# Herausforderung: Verteilung der Schlagwörter



# Validierung der Methode

Verfahren nach Koraljka Golub (2016):

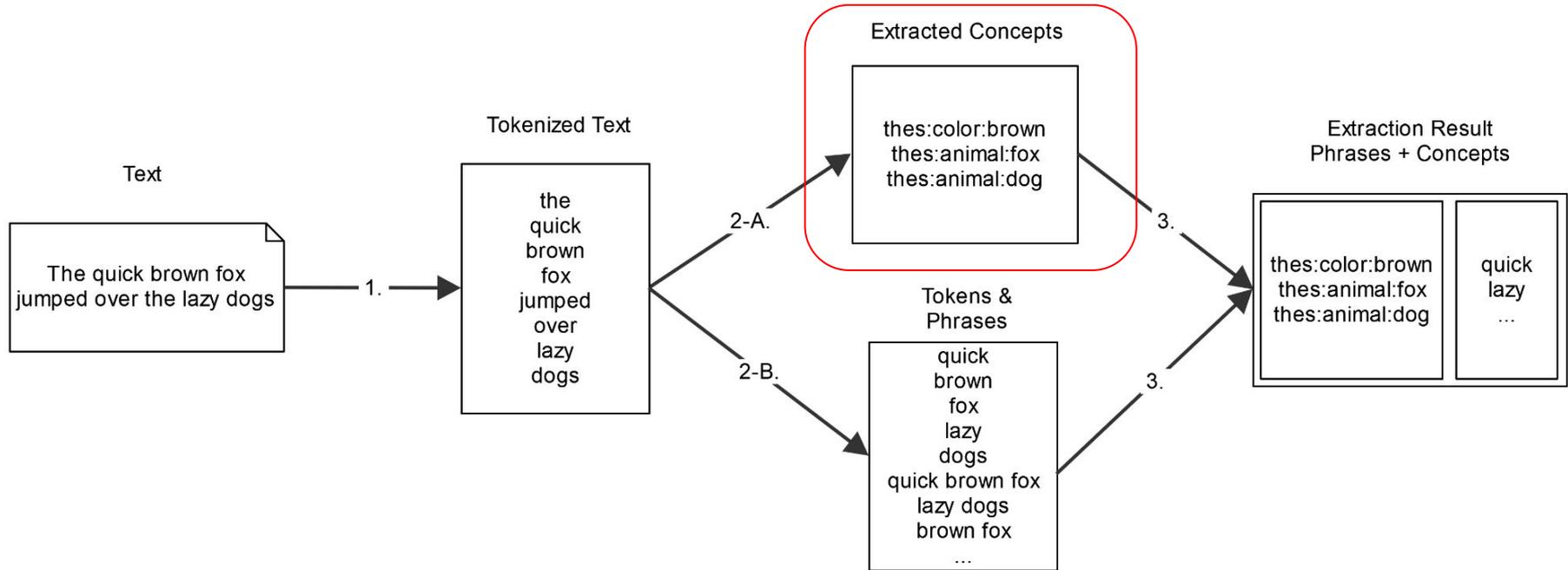
- Bilden eines Goldstandard (annotiertes Referenzkorpus)
- Indexierung des Referenzkorpus mit dem Thesaurus
- Bewertung der extrahierten Metadaten

[Benchmark: F1-Score > 0,70 (PETRUS) ]

# Goldstandard

- 700 Dokumente aus 14 Sachgruppen
- Sachgruppenauswahl angelehnt an PETRUS Auswertungen
  - 004, 020, 100, 150, 320, 330, 340, 370, 510, 530, 610, 620, 630, 780, ~~830~~, ~~900~~
- Herausforderungen:
  - Suchfilter der DNB liefern falsche / irrelevante Ergebnisse
  - Qualität der Metadaten
  - Verteilung der im Volltext verfügbaren Publikationen
  - Fragmentierung der DDC (SG 800, 900, auch 400)
- Manuelle Auswahl der Volltexte anhand der Metadaten der DNB
  - Eindeutige Zuordnung zu einer DDC-Sachgruppe
  - Validierung durch Kuzautopsie

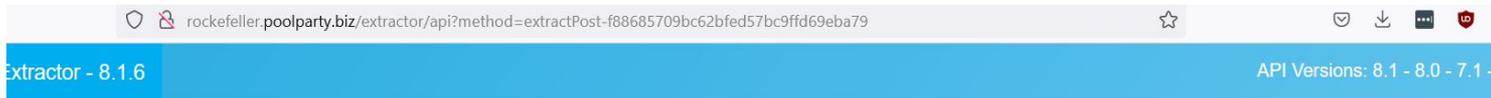
# Funktionsweise des PoolParty Indexers



# Gewichtung und Aggregation

- Gewichtung der Frequenz der Benennung
- Gewichtung der Positionen
  - Die letzte Position des Dokuments besitzt noch  $\frac{1}{4}$  des Positionsgewichts der ersten.
- Gewichtung nach Phrasenlänge:  $\frac{1}{4} (n)$ 
  - Die Benennung „Field programmable gate array“ erhält z.B. das Gewicht 2
- Aggregation der Schlagwörter anhand der übergeordneten Sachgruppen (transitiver Schluss) in Position der obersten Oberbegriffe (skos:topConcept)
  - über API- Parameter gesteuert
  - Polyhierarchie wird bei der Relevanzermittlung berücksichtigt

# API-Funktionen



## POST /api/extract

[text] Extracts and returns meaningful metadata like concepts and terms from a given text

### REQUEST

Request URL: `/extractor/api/extract`

Request Methods: `POST`

Content-Type: `application/x-www-form-urlencoded`

### Request Parameters

Parameter	Type	Required	Description
<code>categorizationWithPpxBoost</code>	boolean	false	Use Extractor boosting, default = false
<code>categorize</code>	boolean	false	Categorization extraction, default = false
<code>conceptMinimumScore</code>	Double	false	Minimum required score of concepts, default = 0
<code>conceptSchemeFilters</code>	Array of String	false	Concept scheme URI filters
<code>corpusScoring</code>	Array of String	false	Corpus term scoring. Enabled if corpusIds (UUID) are provided.
<code>customAttributeFilters</code>	Array of <a href="#">CustomProperty</a>	false	Custom attribute (property uri and string value) filters
<code>customClassFilters</code>	Array of String	false	Custom class URI filters
<code>disambiguate</code>	boolean	false	Use thesaurus based disambiguation, default = false
<code>displayText</code>	boolean	false	Include text extracted from url in response, default = false
<code>documentClassifierIds</code>	Array of String	false	Enable document classification by giving the document classifier IDs as input



# Auswertung

- Bewertung der harten Kategorisierung (Golub)

- Recall
- Precision
- F1 - Score

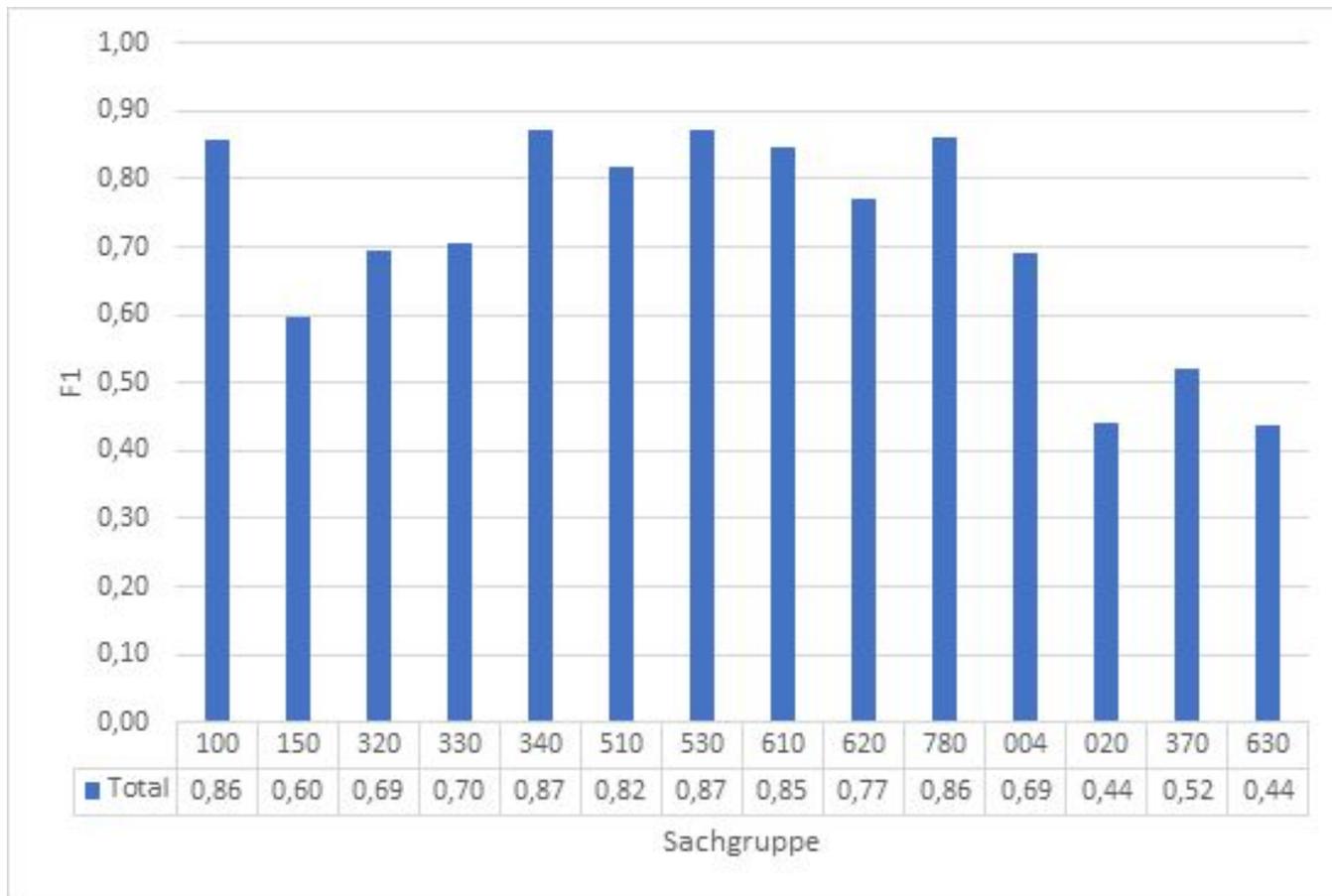
$$F_1 = 2 \frac{\textit{Precision} \times \textit{Recall}}{\textit{Precision} + \textit{Recall}}$$

- Bewertung der rangordnenden Kategorisierung

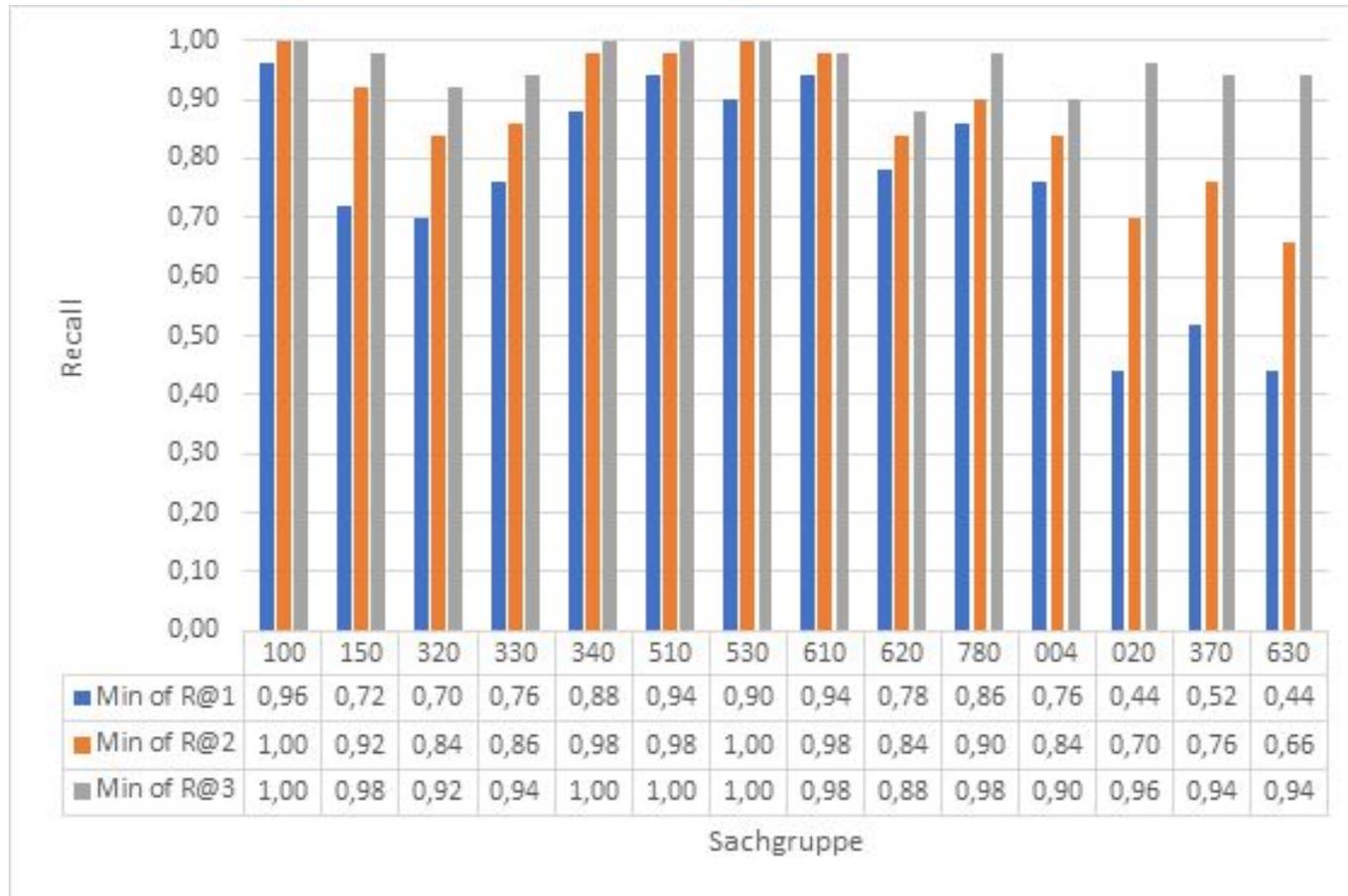
- Recall@k
- Mean Reciprocal Rank

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\textit{Rank} (i)}$$

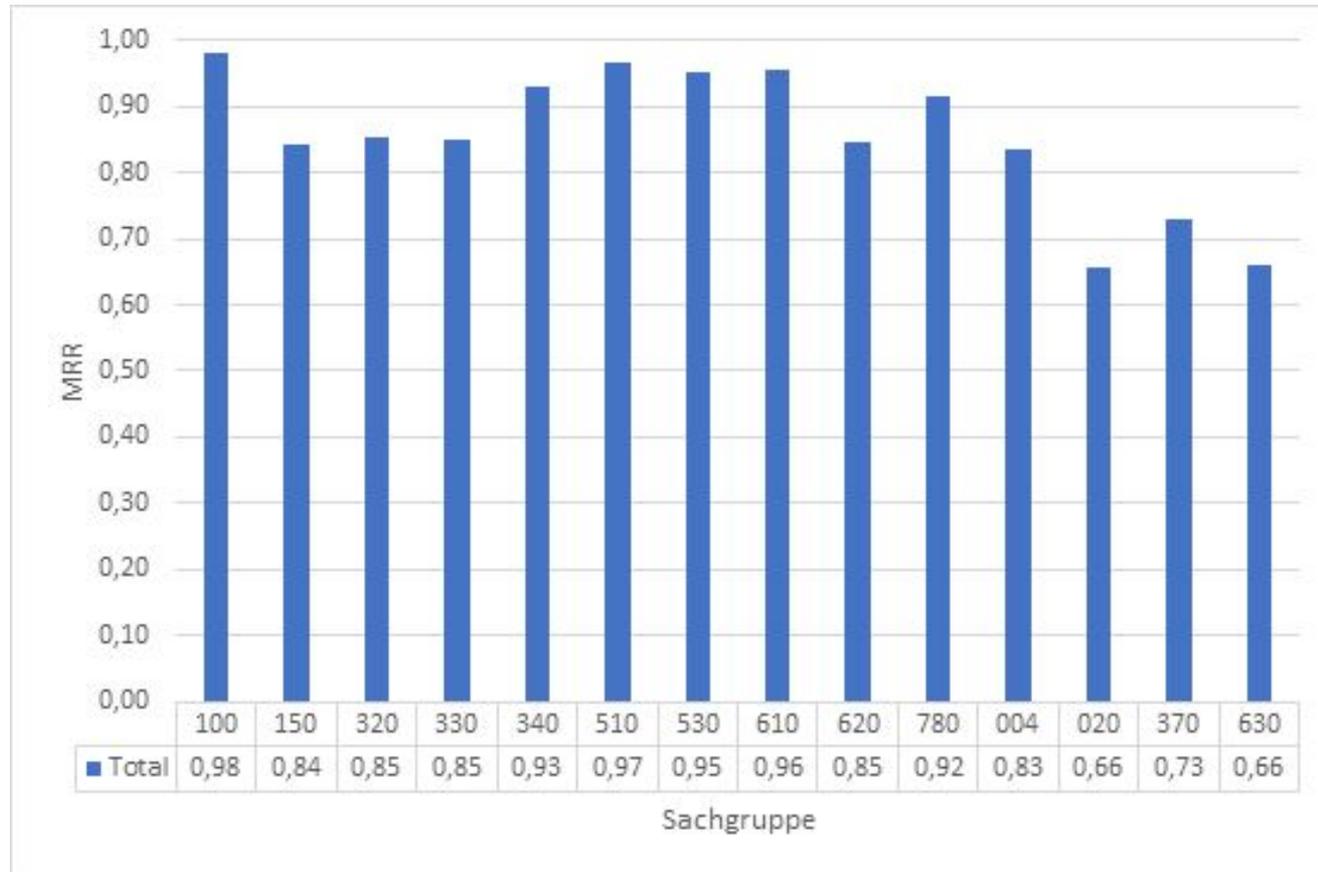
# F1-Score (Median = 0,74)



# Recall der ersten drei Ränge



# Mean Reciprocal Rank (Median = 0,85)



# Wohin gehört die Wirtschaftsinformatik?

	
<b>Link zu diesem Datensatz</b>	<a href="https://d-nb.info/1021499684">https://d-nb.info/1021499684</a>
<b>Art des Inhalts</b>	Hochschulschrift
<b>Titel</b>	Branchenspezifische IT-Innovationssysteme[[Elektronische Ressource]] : von der Analyse zur Intervention ; am Beispiel des IT-Innovationssystems für Krankenhäuser in Deutschland / Paul Drews. Betreuer: Arno Rolf
<b>Person(en)</b>	Drews, Paul (Verfasser) Rolf, Arno (Akademischer Betreuer)
<b>Verlag</b>	Hamburg : Staats- und Universitätsbibliothek Hamburg
<b>Zeitliche Einordnung</b>	Erscheinungsdatum: 2012
<b>Umfang/Format</b>	Online-Ressource
<b>Hochschulschrift</b>	Hamburg, Universität Hamburg, Diss., 2012
<b>Persistent Identifier</b>	URN: <a href="https://nbn-resolving.org/urn:nbn:de:gbv:18-55396">urn:nbn:de:gbv:18-55396</a>
<b>URL</b>	<a href="http://ediss.sub.uni-hamburg.de/volltexte/2012/5539/">http://ediss.sub.uni-hamburg.de/volltexte/2012/5539/</a> (Verlag) (kostenfrei zugänglich)
<b>Sprache(n)</b>	Deutsch (ger)
<b>Schlagwörter</b>	Innovation ; Informationstechnik ; Krankenhaus ; Wirtschaftsinformatik ; Grounded theory
<b>Sachgruppe(n)</b>	360 Soziale Probleme, Sozialdienste, Versicherungen

Sachgruppe Rang 1	Sachgruppe Rang 2	Sachgruppe Rang 3	Relevantes Schlagwort
330 – Wirtschaft: 0.11	650 – Management: 0.10	300 – Soz.: 0.08	Innovation[1] 412

[1] <https://d-nb.info/gnd/4027089-0>

# Zusammenfassung

- Das gewählte Verfahren kategorisiert Texte mithilfe einer Beschreibungslogik und NLP- Technologien
- Die Ergebnisse sind zuverlässig
- Die entstehenden Metadaten können nach dem Regelwerk nachvollzogen und validiert werden
- Die entstehenden Daten sind FAIR
- Das Verfahren eignet sich zur Ergänzung und Verbesserung automatisch und intellektuell erstellter Indexate und zur kollaborativen Erschließung
- Eine umfassende Anwendung erfordert weitere Kuratierung des Schlagwortkatalogs

Vielen Dank für die Aufmerksamkeit!  
Fragen?

[sebastiangabler@gmx.net](mailto:sebastiangabler@gmx.net)

<https://theses.univie.ac.at/detail/60927/>