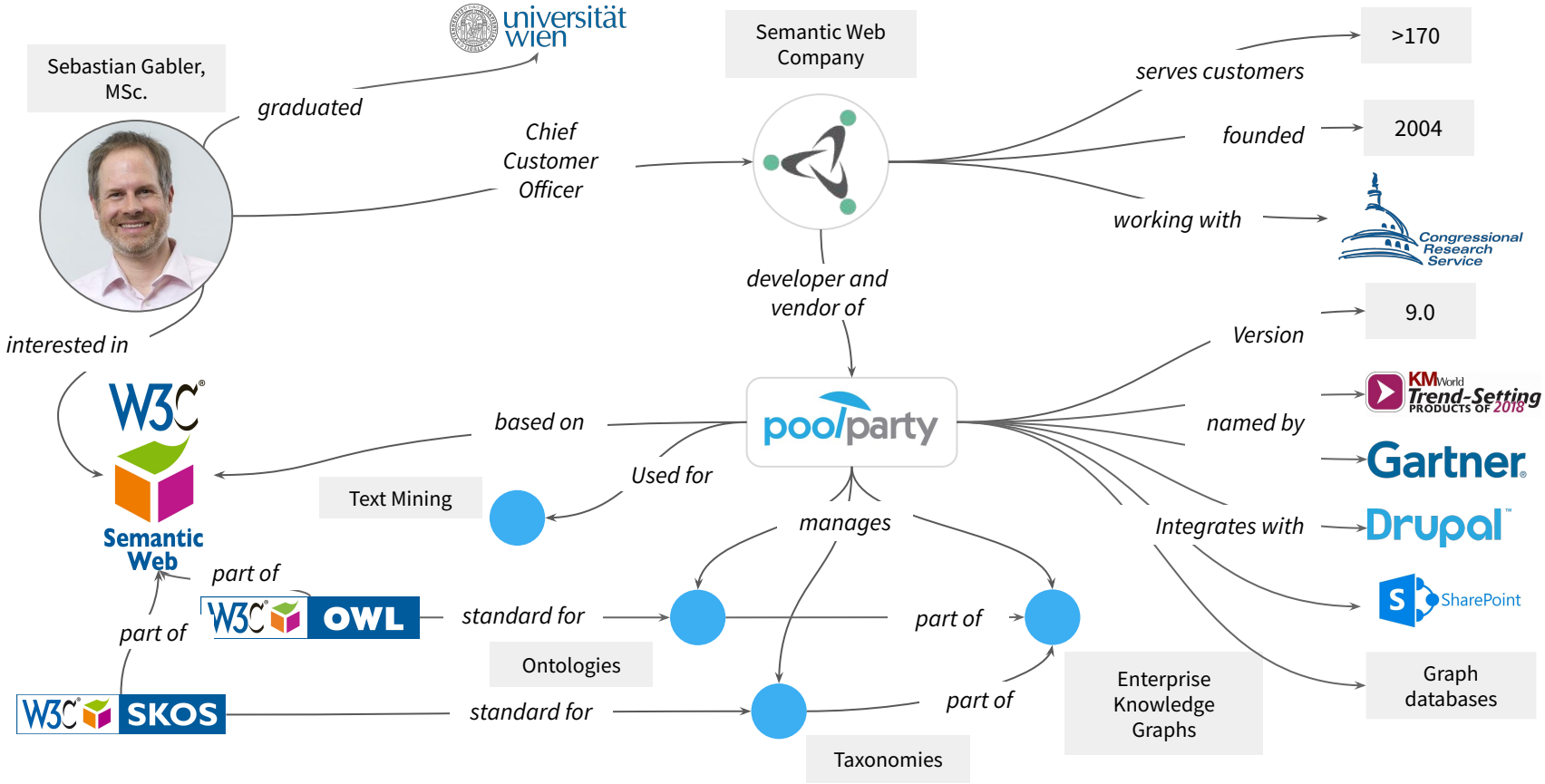
Vergabe von DDC-Sachgruppen mittels eines Schlagwort-Thesaurus

Dipl. Tonm. Sebastian Gabler, MSc.
sebastian.gabler@gmx.net

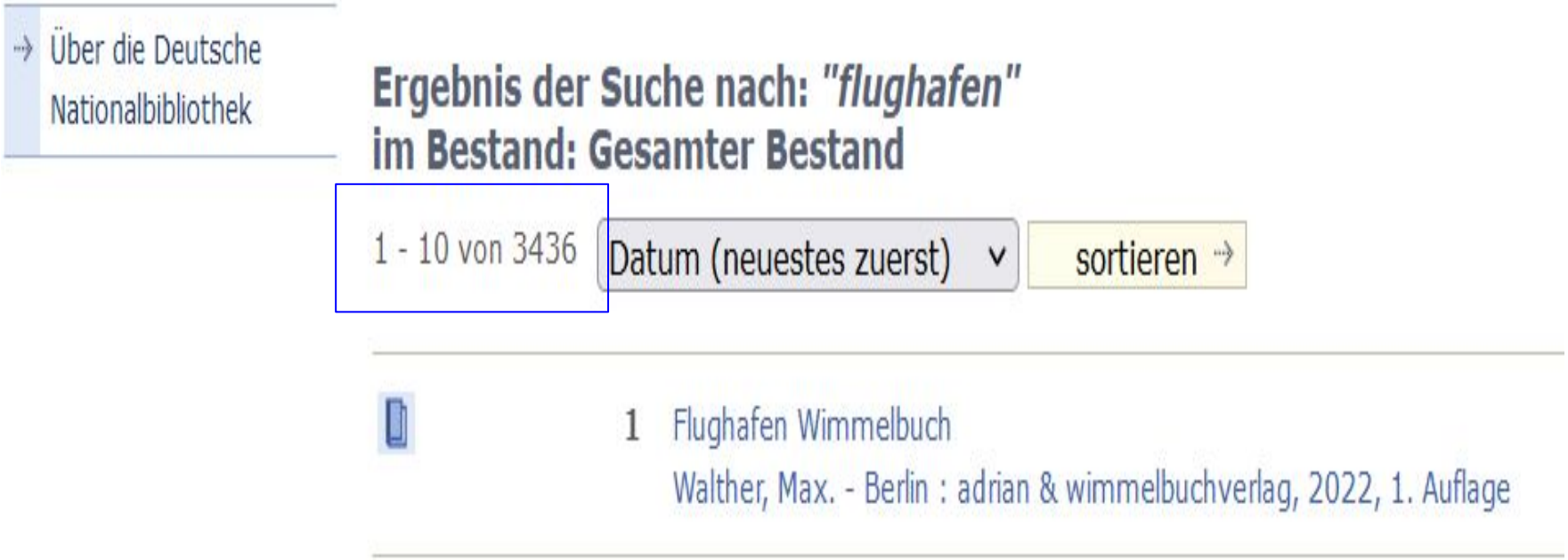

Sebastian Gabler, MSc.
universität wien
graduated
Chief Customer Officer
Semantic Web Company
serves customers
>170
founded
2004
working with
Congressional Research Service
developer and vendor of
Version
9.0
interested in
W3C
Semantic Web
based on
pool party
Used for
Text Mining
named by
KM World Trend-Setting PRODUCTS OF 2018
Gartner
Integrates with
Drupal
SharePoint
manages
part of
W3C OWL
standard for
Ontologies
part of
Enterprise Knowledge Graphs
Graph databases
part of
W3C SKOS
standard for
Taxonomies
part of

Agenda

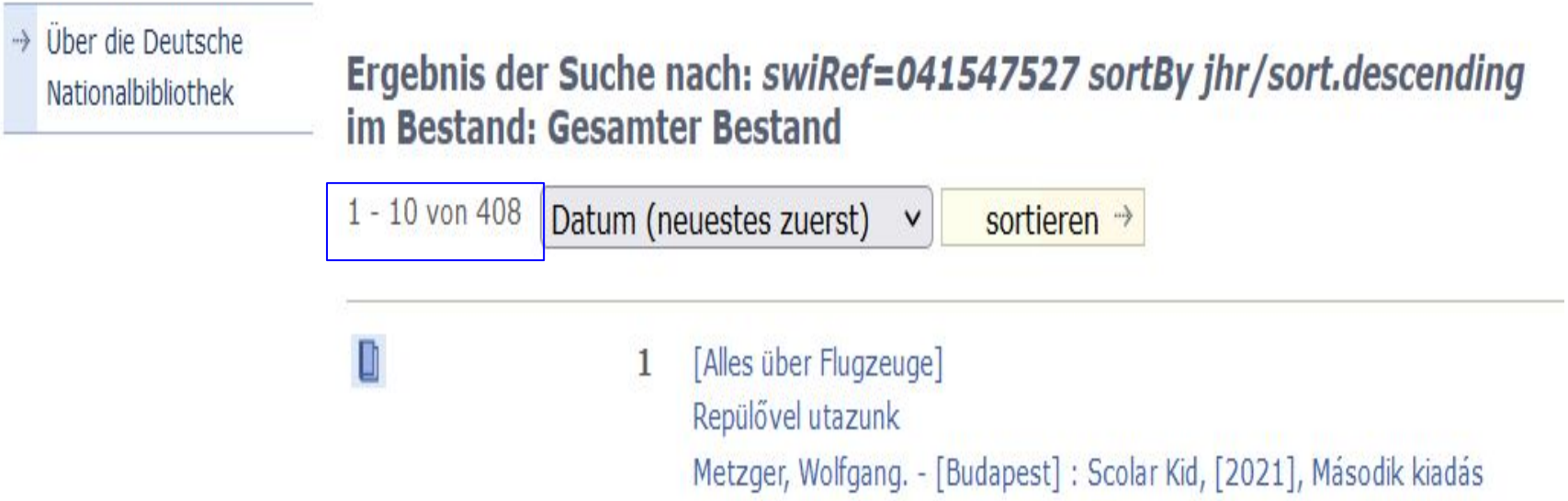
- Warum braucht Information Retrieval konsistente Metadaten?
- Automatisierung unter Nutzung einer Beschreibungslogik
- Klassen- und dokumentenzentrische Herausforderungen
- Evaluierung des Verfahrens
- Zusammenfassung und Ausblick
- Diskussion

“Information Retrieval benötigt konsistente und präzise Metadaten.”

Stichwortsuche führt zu vielen Resultaten

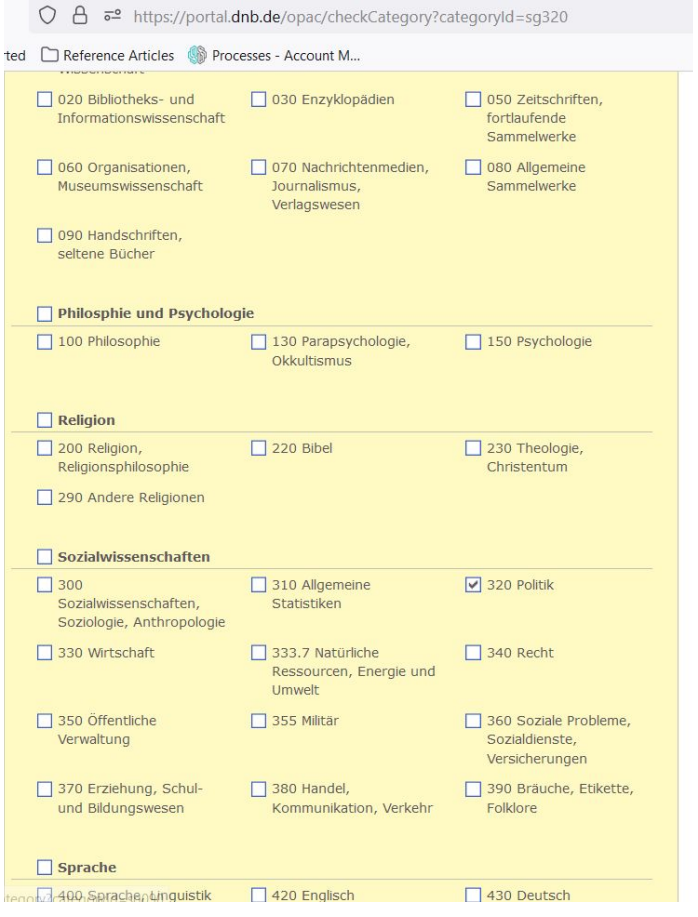


Schlagwortsuche erlaubt hohe Präzision



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, …)



Metasuche und Facettierung über Klassifikation

WebDewey Search — Suche mit der Dewey-Dezimalklassifikation

Suchbegriff oder Notation: SUCHE ☐ 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
	Haupttafeln		
300	Sozialwissenschaften	0 (DNB) 28 (GBV) 0 (HeBIS) 12 (SUB) 204 (SWB) 6098 (FUB)	438935 (DNB) 2174504 (GBV) 2537399 (HeBIS) 508441 (SUB) 1416551 (SWB) 0 (FUB)
380	Handel, Kommunikation & Verkehr	0 (DNB) 3 (GBV) 0 (HeBIS) 1 (SUB) 3 (SWB) 187 (FUB)	20603 (DNB) 133423 (GBV) 156817 (HeBIS) 17127 (SUB) 38608 (SWB) 14014 (FUB)
383-388	Kommunikation und Verkehr	316 (DNB) 2097 (GBV) 97090 (HeBIS) 249 (SUB) 2688 (SWB) k.A. (FUB)	13321 (DNB) 24476 (GBV) 97090 (HeBIS) 3751 (SUB) 19197 (SWB) k.A. (FUB)
387	Schifffahrt, Luft-, Weltraumverkehr	12 (DNB) 176 (GBV) 0 (HeBIS) 27 (SUB) 140 (SWB) 54 (FUB)	2784 (DNB) 5937 (GBV) 9989 (HeBIS) 668 (SUB) 3280 (SWB) 812 (FUB)
387.7	Luftverkehr	60 (DNB) 221 (GBV) 0 (HeBIS) 24 (SUB) 260 (SWB) 63 (FUB)	1027 (DNB) 1851 (GBV) 4194 (HeBIS) 214 (SUB) 1450 (SWB) 351 (FUB)
387.73	Luftfahrzeuge und Einrichtungen	7 (DNB) 11 (GBV) 0 (HeBIS) 1 (SUB) 29 (SWB) 2 (FUB)	470 (DNB) 471 (GBV) 1703 (HeBIS) 50 (SUB) 363 (SWB) 44 (FUB)
387.736	**Flughäfen**	**41 (DNB) 106 (GBV) 0 (HeBIS) 10 (SUB) 133 (SWB) 18 (FUB)**	**240 (DNB) 377 (GBV) 589 (HeBIS) 45 (SUB) 276 (SWB) 42 (FUB)**
387.7362	Einrichtungen	16 (DNB) 19 (GBV) 0 (HeBIS) 0 (SUB) 11 (SWB) 0 (FUB)	18 (DNB) 23 (GBV) 71 (HeBIS) 0 (SUB) 14 (SWB) 0 (FUB)
387.7364	Teilbereiche und Dienste	17 (DNB) 28 (GBV) 0 (HeBIS) 2 (SUB) 24 (SWB) 4 (FUB)	40 (DNB) 43 (GBV) 35 (HeBIS) 5 (SUB) 30 (SWB) 6 (FUB)

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

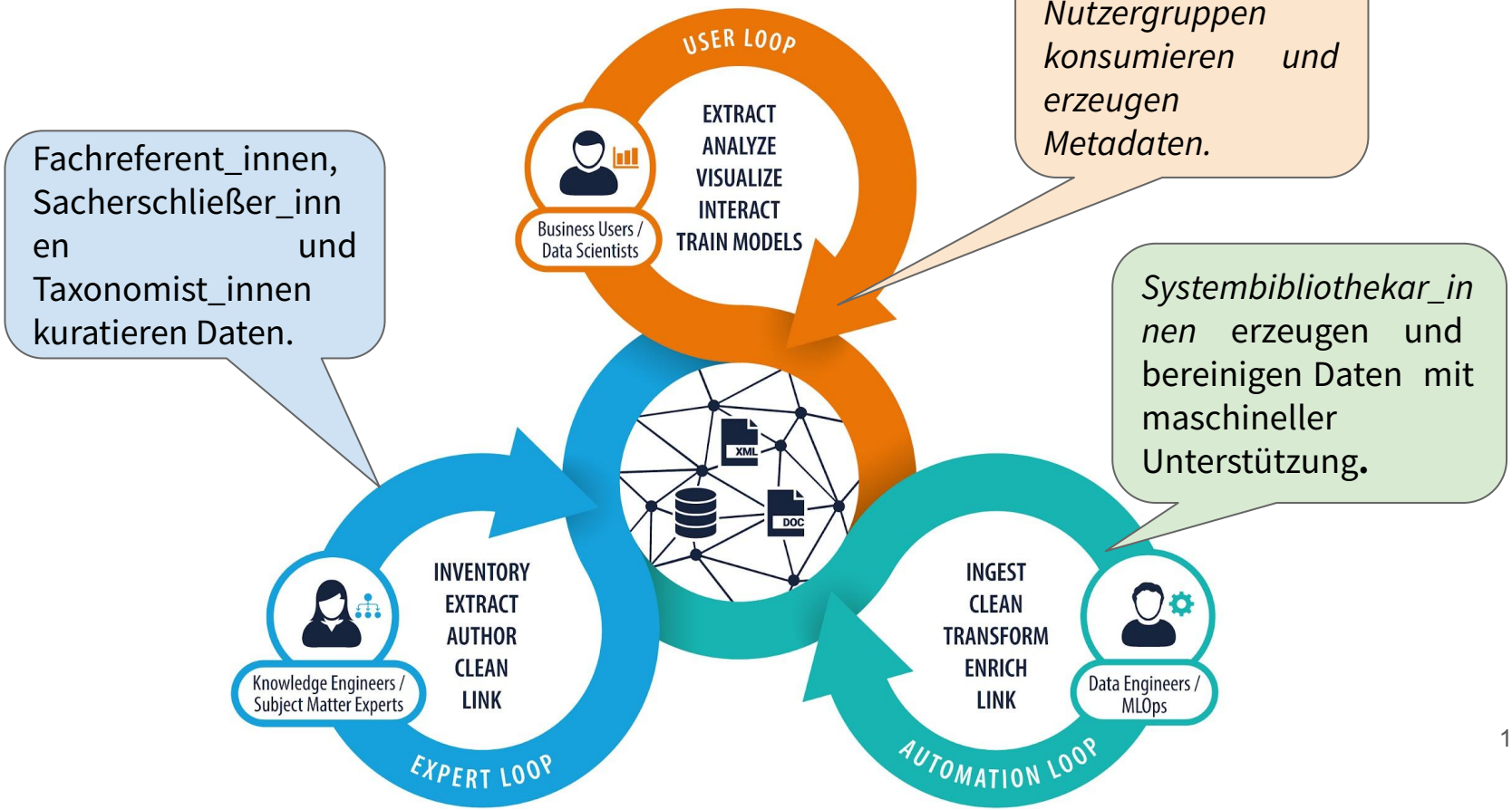
“Sacherschließung benötigt agile Prozesse.”

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



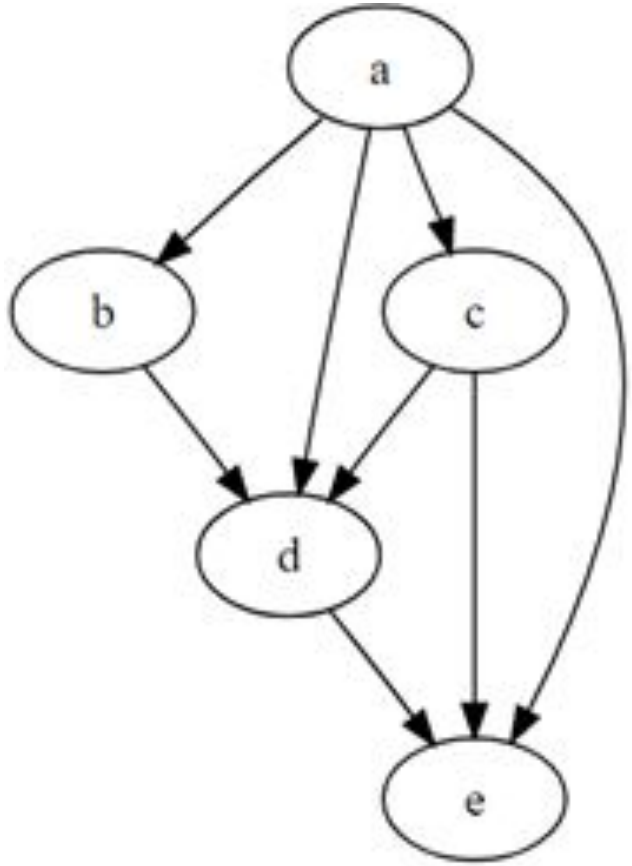
Fachreferent_innen, Sacherschließer_innen und Taxonomist_innen kuratieren Daten.

Verschiedene Nutzergruppen konsumieren und erzeugen Metadaten.

Systembibliothekar_innen erzeugen und bereinigen Daten mit maschineller Unterstützung.

“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:

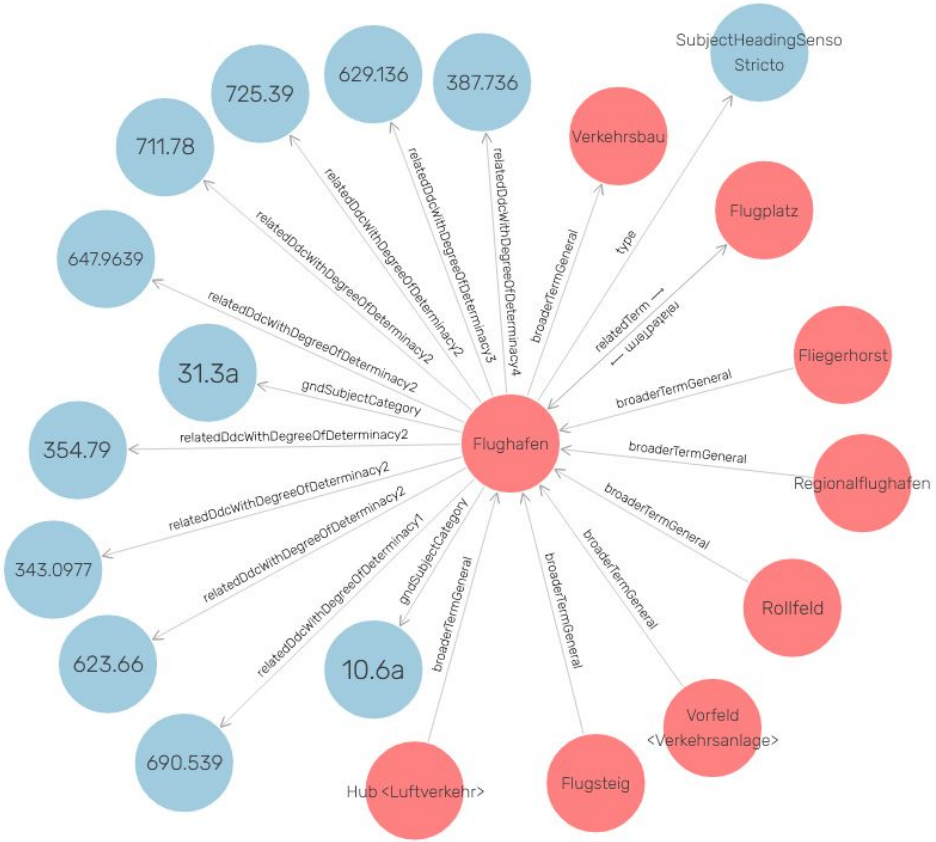
a ist Oberbegriff von b, ist Oberbegriff von c, ist Oberbegriff von e.

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

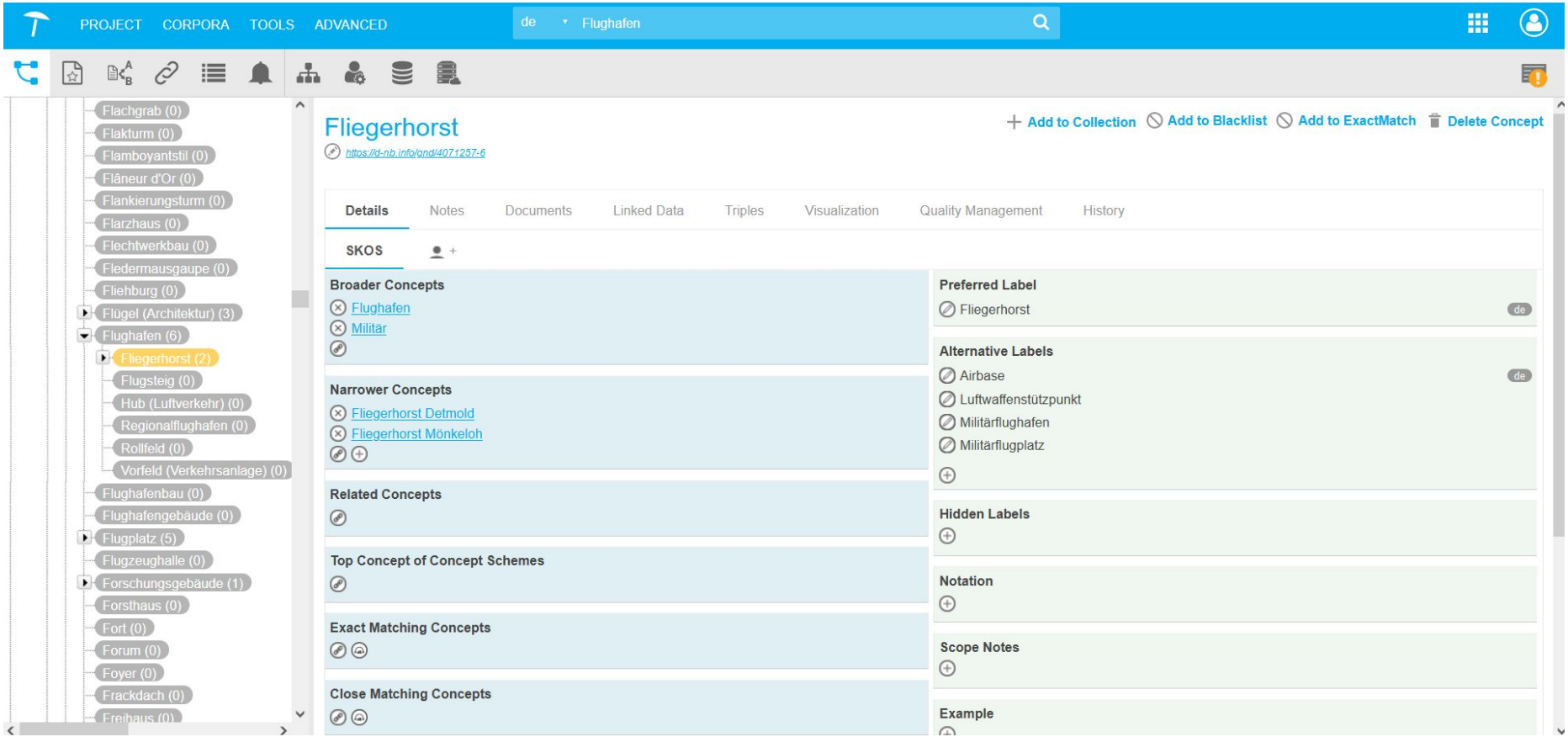
https://d-nb.info/standards/elementset/gnd#gndIdentifier
4154752-4

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

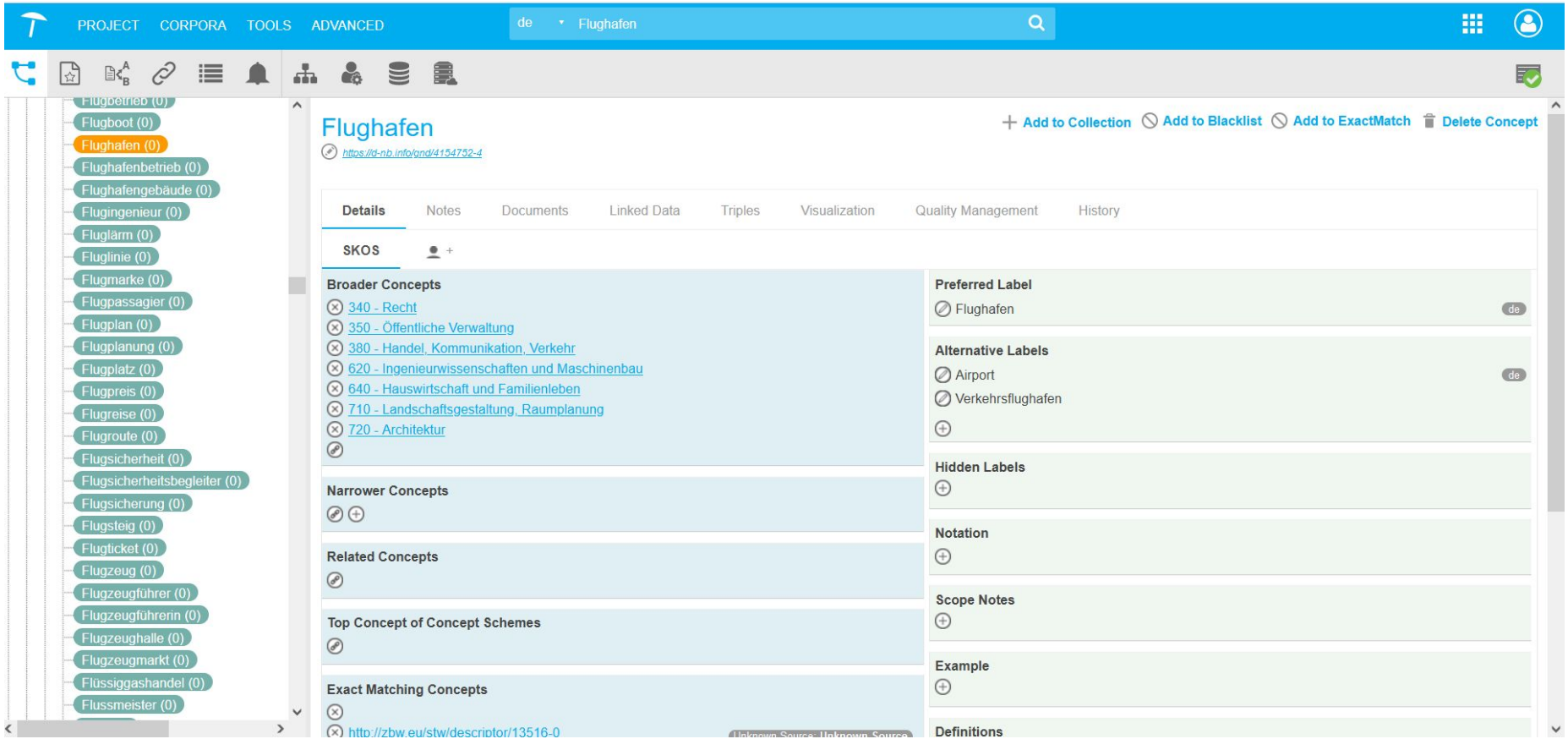
https://d-nb.info/standards/elementset/gnd#preferredNameForTheSubjectHeading
Flughafen

https://d-nb.info/standards/elementset

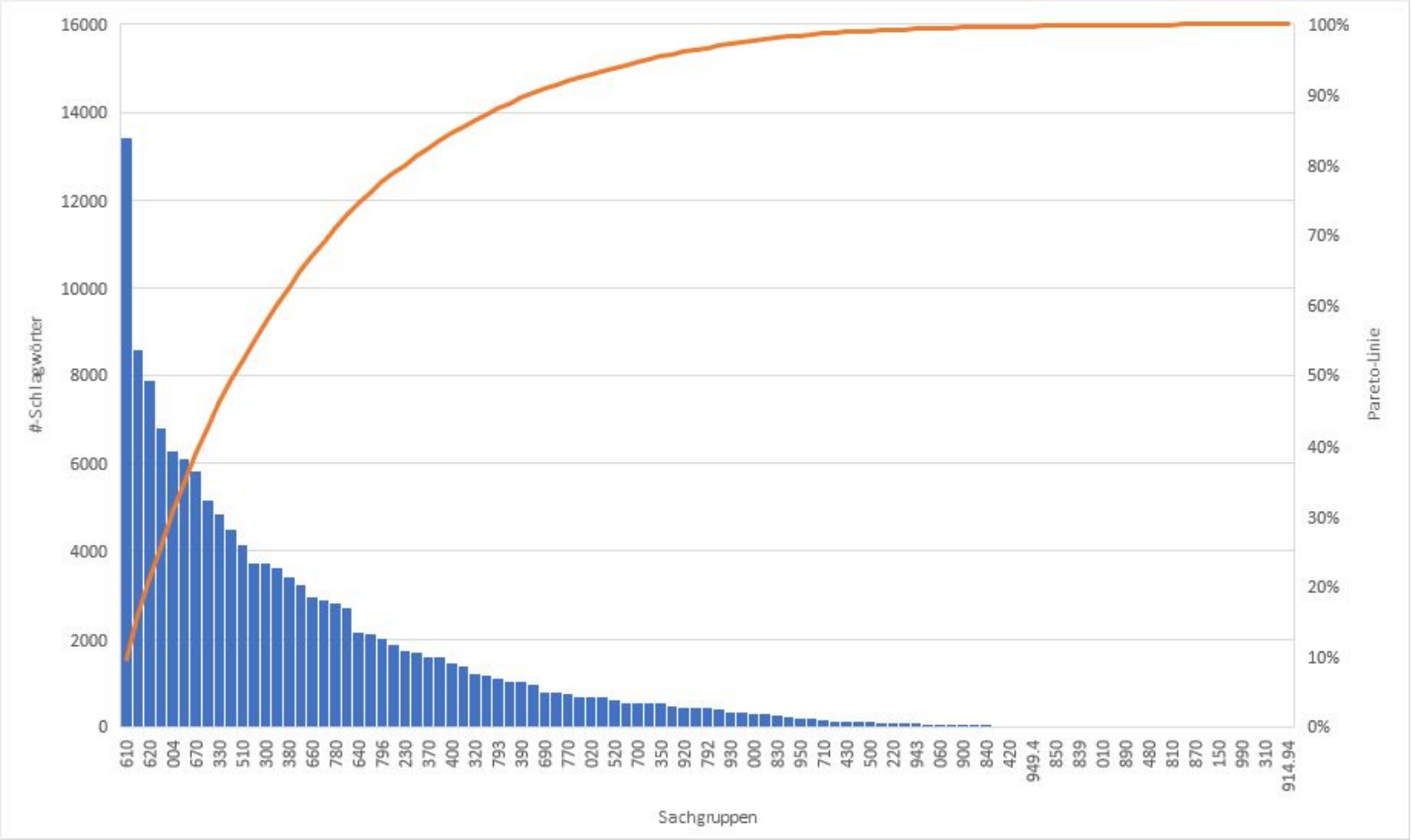
Falsche Polyhierarchie / GND Systematik



Lösung:DDC Sachgruppen als upper ontology



Herausforderung: Verteilung der Schlagwörter



Validierung der Methode

Verfahren nach Koraljka Golub (2016):

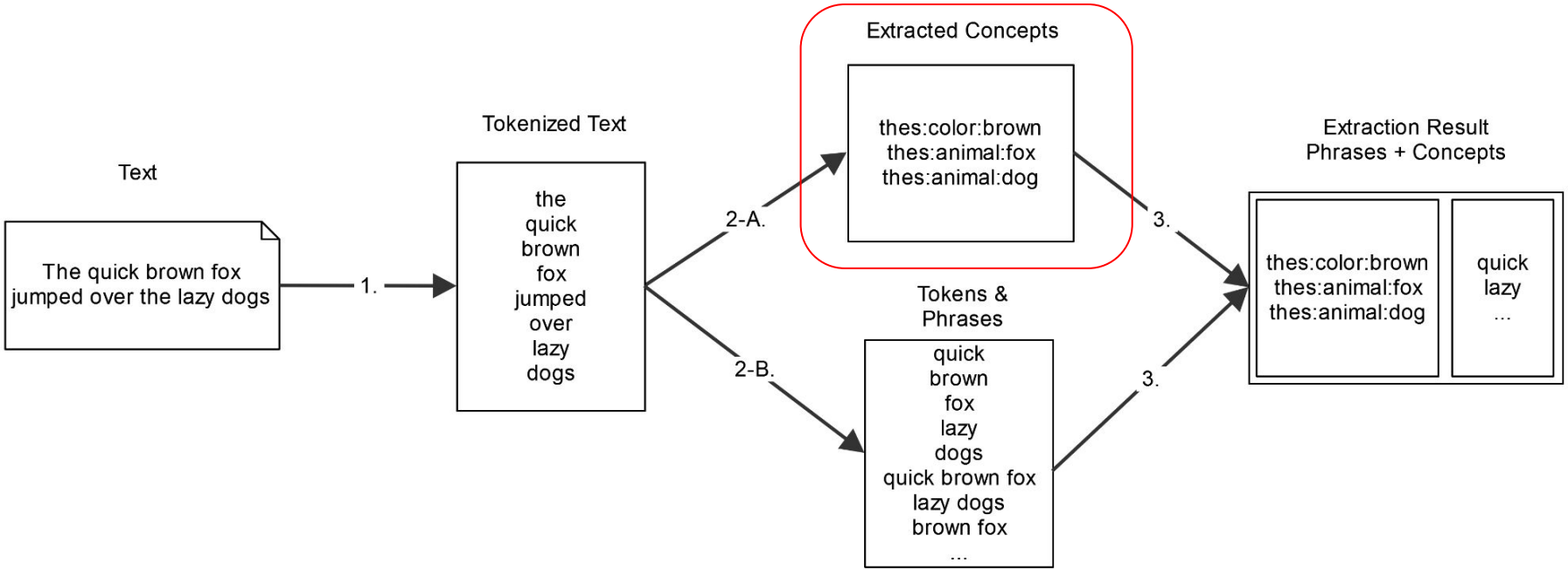
- Bilden eines Goldstandard (annotiertes Referenzkorpus)
- Indexierung des Referenzkorpus mit dem Thesaurus
- Bewertung der extrahierten Metadaten

[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

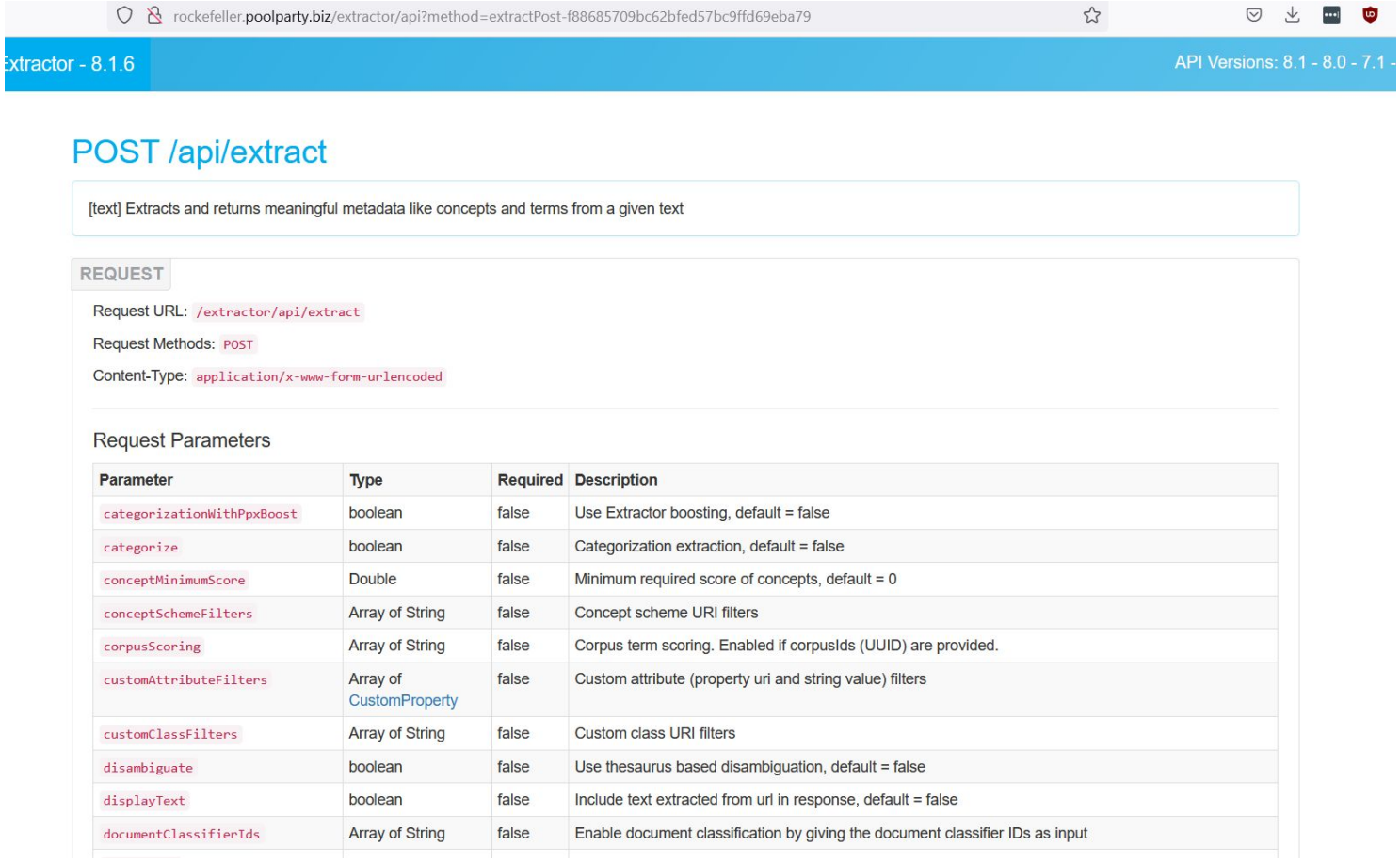
Funktionsweise des PoolParty Indexers



Gewichtung und Aggregierung

- Gewichtung der Frequenz der Benennung
- Gewichtung der Positionen
  - Die letzte Position des Dokuments besitzt noch ¼ des Positionsgewichts der ersten.
- Gewichtung nach Phrasenlänge: ¼ (n)
  - Die Benennung „Field programmable gate array“ erhält z.B. das Gewicht 2
- Aggregierung 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



Indexierung

Link zu diesem Datensatz	https://d-nb.info/999901796
Art des Inhalts	Hochschulschrift
Titel	Die Regionalisierung des österreichischen Wahlverhaltens Bedeutungslosigkeit der Regionalwahlkreise / vorgelegt v
Person(en)	Jenny, Marcelo (Verfasser)
Sekundärausgabe	Online-Ausg.: [2010]. Online-Ressource
Zeitliche Einordnung	Erscheinungsdatum: 2009
Umfang/Format	ca. 2,4 MB
Andere Ausgabe(n)	Druckausg.: Jenny, Marcelo: Die Regionalisierung des öst
Hochschulschrift	Mannheim, Univ., Diss., 2009
Persistent Identifier	URN: urn:nbn:de:bsz:180-madoc-26098
URL	http://madoc.bib.uni-mannheim.de/madoc/volltexte/20 zugänglich) http://madoc.bib.uni-mannheim.de/madoc/volltexte/20
Sprache(n)	Deutsch (ger)
Schlagwörter	Wahlverhalten ; Wahlsystem ; Wahlrechtsreform ; Öster
DDC-Notation	324.9436 [DDC22ger]
Sachgruppe(n)	320 Politik

Online-Zugriff	Archivobjekt öffnen


```
{
  ▾ document: {
      ▸ concepts: […],
      ▾ categories: [
          ▾ {
              prefLabel: "320 - Politik",
              uri: http://dewey.info/class/320/,
              score: 0.17058011049723756,
            ▸ categoryConceptResults: […]
          },
          ▾ {
              prefLabel: "300 - Sozialwissenschaften, Soziologie, Anthropologie",
              uri: http://dewey.info/class/300/,
              score: 0.12810773480662985,
            ▸ categoryConceptResults: […]
          },
          ▾ {
              prefLabel: "510 - Mathematik",
              uri: http://dewey.info/class/510/,
              score: 0.06975138121546962,
            ▾ categoryConceptResults: [
                ▾ {
                    uri: https://d-nb.info/gnd/4038613-2,
                    prefLabel: "Menge",
                    score: 2.0
                },
                ▾ {
                    uri: https://d-nb.info/gnd/4300606-1,
                    prefLabel: "Schiefe Wahrscheinlichkeitsverteilung",
                    score: 1.0
                },
                ▾ {
                    uri: https://d-nb.info/gnd/4078859-3,
                    prefLabel: "Versuchsplanung",
                    score: 1.0
                },
                ▾ {
                    uri: https://d-nb.info/gnd/4137007-7,
                    prefLabel: "Wahrscheinlichkeit",
                    score: 3.0
                },
                ▾ {
                    uri: https://d-nb.info/gnd/4056995-0,
                    prefLabel: "Statistik",
```

Auswertung

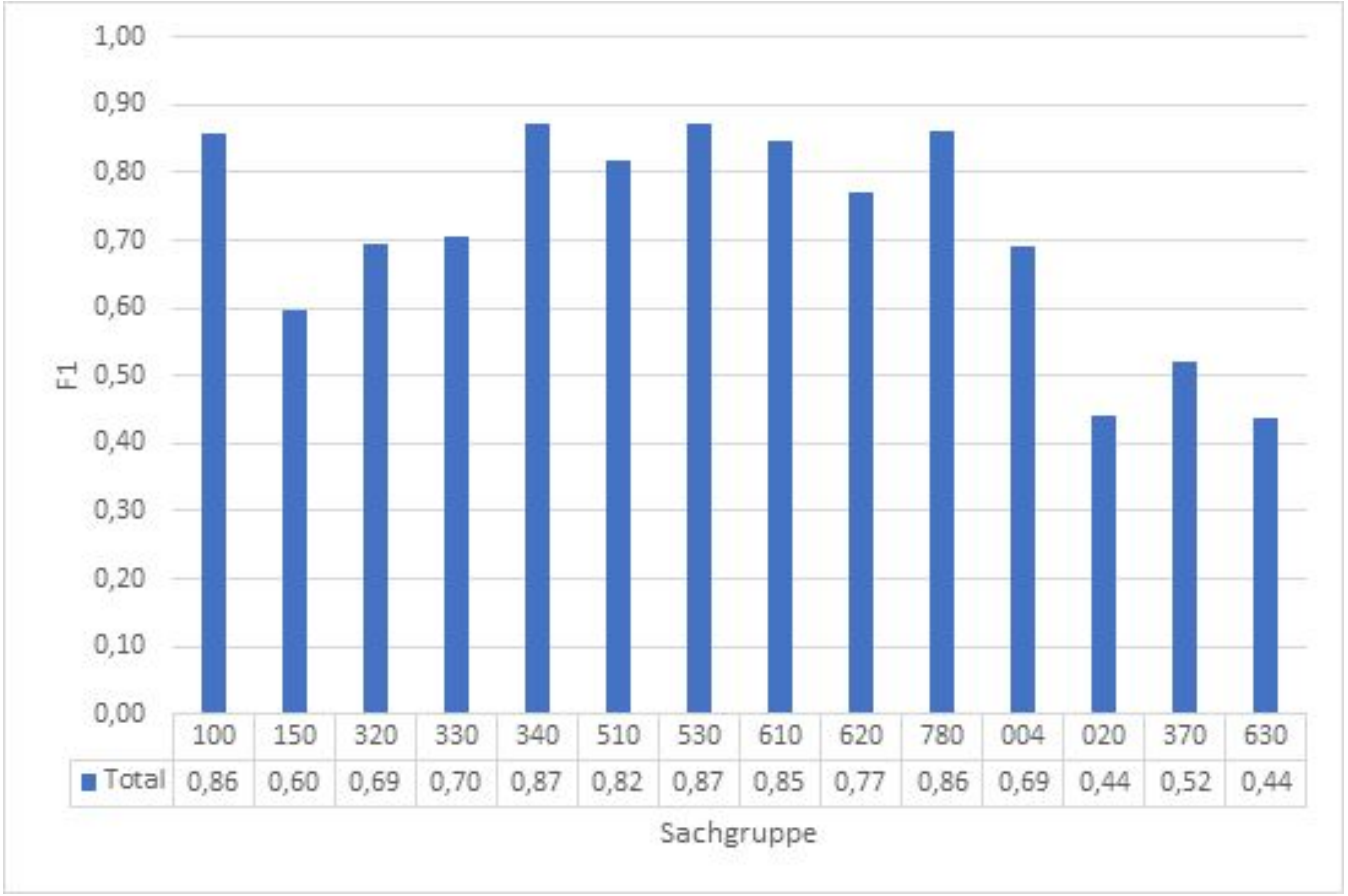
- Bewertung der harten Kategorisierung (Golub)
  - Recall
  - Precision
  - F1 - Score

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

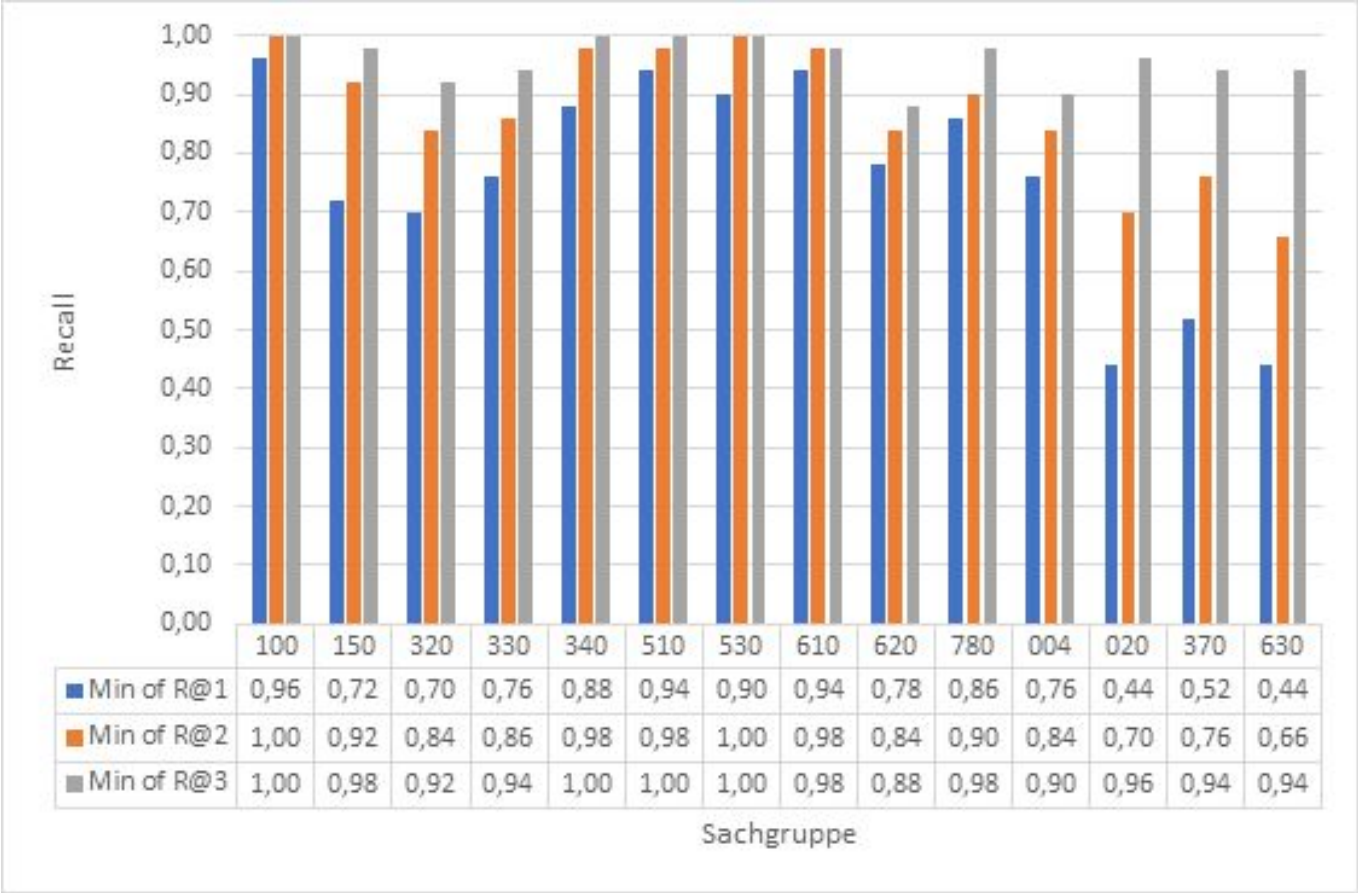
- Bewertung der rangordnenden Kategorisierung
  - Recall@k
  - Mean Reciprocal Rank

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

F1-Score (Median = 0,74)

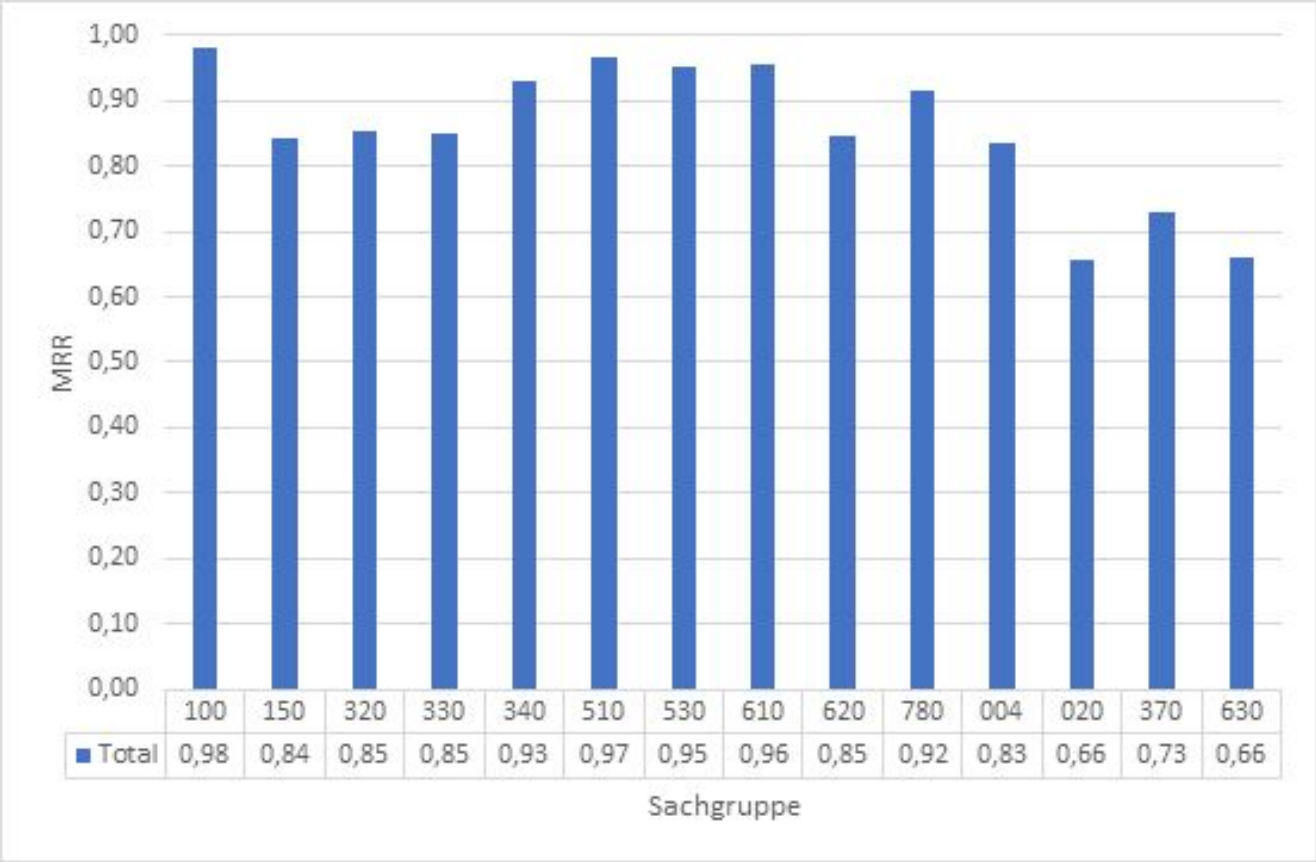


Recall der ersten drei Ränge



	100	150	320	330	340	510	530	610	620	780	004	020	370	630
Min of R@1	0,96	0,72	0,70	0,76	0,88	0,94	0,90	0,94	0,78	0,86	0,76	0,44	0,52	0,44
Min of R@2	1,00	0,92	0,84	0,86	0,98	0,98	1,00	0,98	0,84	0,90	0,84	0,70	0,76	0,66
Min of R@3	1,00	0,98	0,92	0,94	1,00	1,00	1,00	0,98	0,88	0,98	0,90	0,96	0,94	0,94

Mean Reciprocal Rank (Median = 0,85)



Wohin gehört die Wirtschaftsinformatik?

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

Vielen Dank für die Aufmerksamkeit!
Fragen?

sebastiangabler@gmx.net

https://utheses.univie.ac.at/detail/60927/