

coli-ana

**Automatische Analyse der Dewey Dezimal Klassifikation
Eine Dienstleistung der Verbundzentrale des GBV**

Uma Balakrishnan/Jakob Voß



Hochschule der Medien Stuttgart, 21-06-2022





CC-BY-SA Jônatas Cunha <https://w.wiki/4SBh>

Agenda

- Projekt Colibri
- Dewey Dezimal Klassifikation (DDC)
- Herausforderungen
- Workflow
- Use cases
- Technische Umsetzung



CC-BY-SA Jônatas Cunha <https://w.wiki/4SBh>

Initiiert
von Dr.
Ulrike
Reiner
(2003)

Projekt Colibri

Gefördert
durch die
VZG

Context Generator & Linguistic Tools for Bibliographic Retrieval Interface



CC-BY-SA Jônatas Cunha <https://w.wiki/4SBh>

Colibri Research Questions

Is it possible to...

Q1. ...classify automatically bibliographic title records using DDC?

Q2. ...analyse automatically molecular DDC notations
into atomic DDC notations?

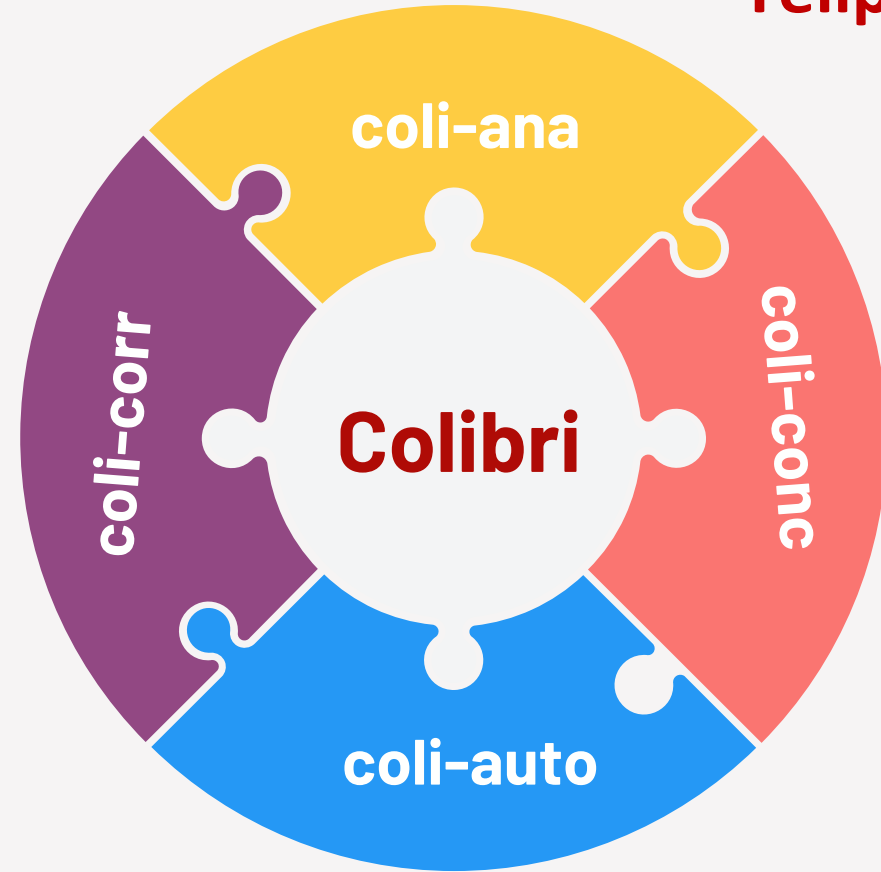
Q3. ...improve automatic classification & retrieval
by means of atomic DDC notations?

atomic DDC notation: a semantically indecomposable notation that represents a DDC class

molecular DDC notation: a notation that is syntactically decomposable into atomic DDC notations

DDC notation: dno

Teilprojekte



- coli-ana: automatic analysis
- coli-conc: concordances
- coli-corr: correctness
- coli-auto: automatic classification

DDC

- **Aktiv in Anwendung über 100 Jahre**
- **Große User Community** weltweit
- VZG Mitglied des **Dewey Consortium** in 2000
- **Projekt DDC Deutsch** 2002
- European DDC User Group seit 2007
- **Dynamisches System**
- **Fine strukturierte Notationen**
- **Ständiger Zuwachs an** DDC Notationen im K10plus Katalog aus externen Quellen
- **Mindestens 1 Mio. einzigartige DDC** gebildete Notationen im K10plus Katalog



Main Classes	
000	Computer science, information & general works
100	Philosophy & psychology
200	Religion
300	Social sciences
400	Language
500	Science
600	Technology
700	Arts & recreation
800	Literature
900	History & geography

DDC System und Notationen

WebDewey

SEARCH

DDC 23											
Main Classes	000	100	200	300	400	500	600	700	800	900	
Tables	T1	T2	T3	T3A	T3B	T3C	T4	T5	T6		
Manual	Introduction	Glossary	Relocations & Discontinuations								

Abridged Edition 15											
Main Classes	000	100	200	300	400	500	600	700	800	900	
Tables	T1	T2	T3	T4							
Manual	Introduction	Glossary	Relocations & Discontinuations								

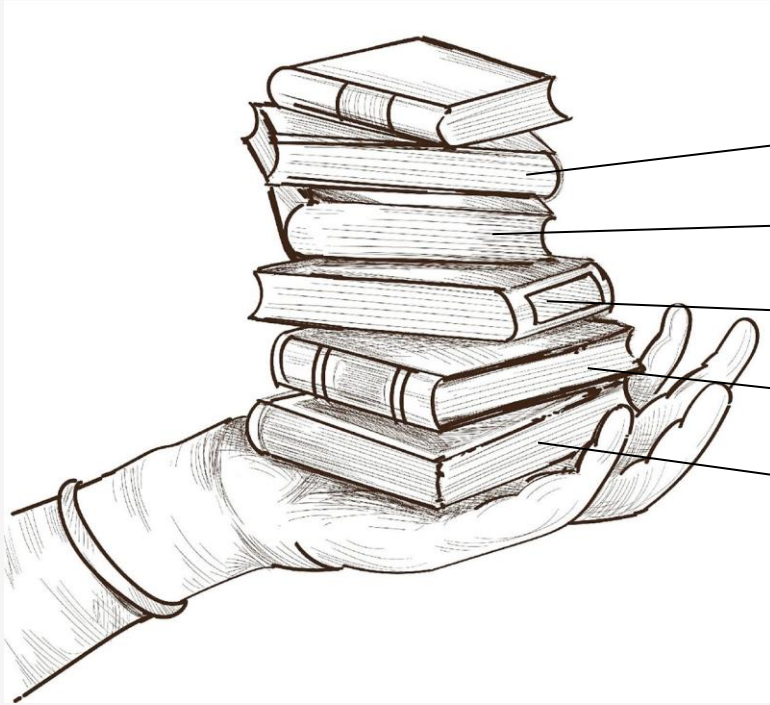
Main Classes

100	Philosophy & psychology
100	Philosophy
110	Metaphysics
120	Epistemology
130	Parapsychology & occultism
140	Philosophical schools of thought
150	Psychology
160	Philosophical logic
170	Ethics
180-190	History, geographic treatment, biography

Main Classes

	Philosophy & psychology
	Philosophy
100	Philosophy, parapsychology and occultism, psychology
101	Theory of philosophy
102	Miscellany of philosophy
103	Dictionaries, encyclopedias, concordances of philosophy
[104]	[Unassigned]
105	Serial publications of philosophy
106	Organizations and management of philosophy
107	Education, research, related topics of philosophy
108	Groups of people
109	History and collected biography

Komplexität der DDC-Notationen



331.892829225209712743090511

700.9044074747

754.09109033

700.23

700

DDC-Notationssynthese

Zusammengesetzt **nach komplexen Regeln** aus mehreren Teilen der DDC-Notationen

Create built number: ⓘ 666.4

[666.4](#) Pottery materials, equipment, processes

Add to base number [666.4](#) the numbers following 738.1 in 738.12-738.15, e.g., kilns [666.436](#)

ADD

EDIT LOCAL

CANCEL

Synthesized number components ⓘ 666.444

[666.4](#) Pottery materials, equipment, processes

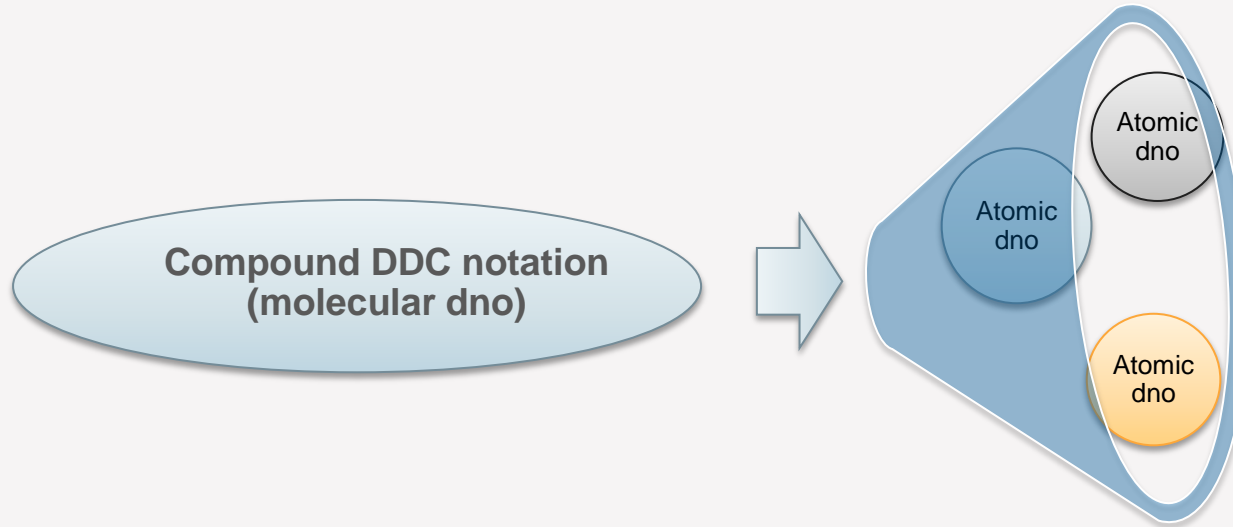
+ [738.144](#) Glazing

Pottery glazing technology

666.4 + 738.144 ⇒ 666.444

coli-ana

Tool für automatische Analyse der zusammengesetzten DDC-Notationen



Ziele

Verbesserung des Retrievals

- Unterstützung von Suchmöglichkeiten (atomic dnos)
Beispiel: T1--09044 (DDf *1940-1949)
Ergebnis: alle Titel , die mit dem Zeitraum 1940-1949 verbunden sind
- Erweiterung der Volltextsuche mittels Bennennungen
- Auswertung und Ranking von ähnlichen Publikationen

Analyse & Verbesserung der Inhaltserschließung

Beispiel: DDC 700 "Arts" and T1--0901-0905:074 "Museums, collections, exhibitions"
=> BK 20.13 "Art exhibition".

Verbesserung der Darstellung im K10plus Verbundkatalog

K10plusPPN: 374538816

Titel: 1803 - die gelehrten Mönche und das Ende einer 1000-jährigen Tradition : "In den Klöstern gediehen die größten Männer"; [Ausstellung vom 28. Mai bis 24. August 2003 Museum Obermünster, Regensburg / [hrsg. vom Bischöflichen Ordinariat Regensburg. Schriftleitung: Maria Baumann]

Beteiligt: Baumann, Maria

Körperschaft: Diözesanmuseum Regensburg
Diözese (Regensburg)

Erschienen: Regensburg : Schnell und Upmann, 2003

Umfang: 71 S. : Ill.

Sprache(n): Deutsch

Schriftenreihe: Kataloge und Schriften / Diözesanmuseum Regensburg

ISBN: 3-7954-1587-X

Sonstige Nummern: OCoLC: 163150381 →

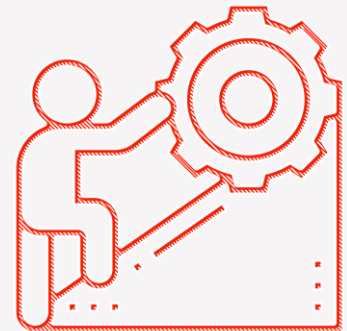
RVK-Notation: G:bn S:rg | BO 1268 | LH 47480 | NS 5500 | NS 1925 → Ähnliche Literatur

Sachgebiete: Basisklassifikation: 11.54 (Katholizismus)
SSG-Nummer(n): 8,1

Tooltip:
Benennung der RVK-Notation "BO 1268"
Zur Navigation in RVK Online klicken Sie bitte auf den INFO-Button
BO 1268
Theologie und Religionswissenschaften
/ Patrologie und Kirchengeschichte
/ Hand- und Lehrbücher sowie Gesamtdarstellungen
/ Bayern
/ Bayern nach Orten (CSN des Namens)
/ Regensburg (Stadt)

Herausforderungen

- **Umfangreiches System mit über 51.700** Klassen
- **Kompliziertes Verfahren der Notionssynthese**
 - Haupttafeln, Hilfstafeln, interne Tabellen, Standardschlüssel, Anmerkungen, Regeln
 - **Über 9.987 Regeln** (Beispiele: add note, authorisation, discontinuation, include note, class here, class elsewhere, revision,...)
 - Gebündelt in 60 Rule parts
- **Möglichkeit sehr feine und themenspezifische** Notationen zu bilden



Beispiel: Komplette Analyse einer synthetisierten DDC-Notation

700.90440747471

7-----

70-----

700-----

700.9-----

700.9-----

700.904-----

-0-----

--0-----

--0.9-----

--0.904-----

--0.904-----

--0.9044-----

---.---07----

---.---074----

---.---7----

---.---7----

---.---74--

---.---74--

---.---747-

---.---747-

---.---7471

Arts & recreation (700)

Arts (700)

The arts (700)

Standard subdivisions of the arts (700.1-700.9)

History, geographic treatment, biography of the arts (700.9)

Arts--20th century,... (700.904)

facet indicator (0)

Table 1. Standard Subdivisions (T1--0)

History, geographic treatment, biography (T1--09)

Historical periods (T1--0901-0905)

*20th century, 1900-1999 (T1--0904)

*1940-1949 (T1--09044)

Museums, collections, exhibits; collecting objects (T1--0901-0905:07)

Museums, collections, exhibits (T1--0901-0905:074)

Modern world; extraterrestrial worlds (T2--4-9)

North America (T2--7)

Specific states of United States (T2--74-79)

Northeastern United States (New England and Middle Atlantic states) (T2--74)

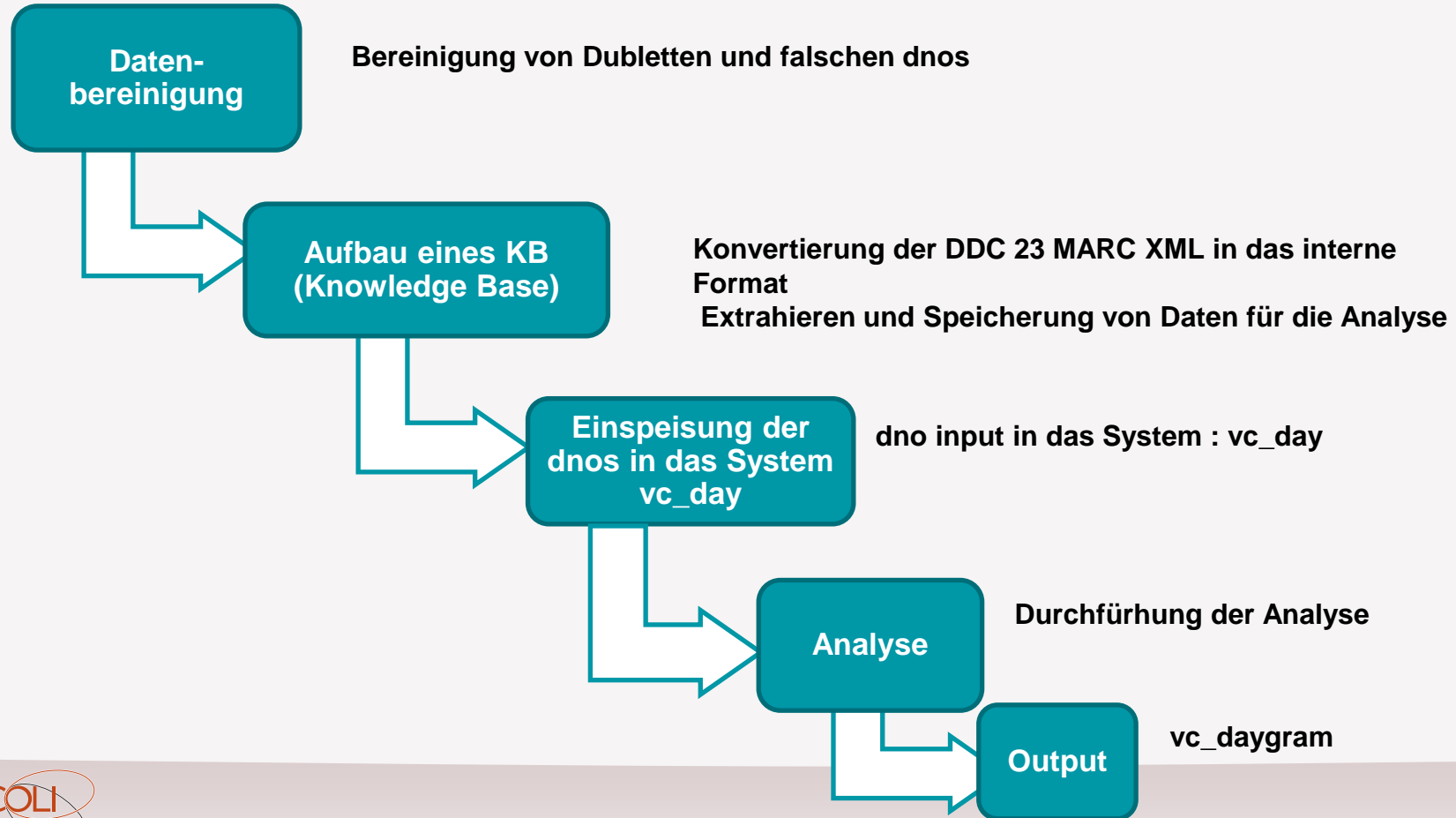
Middle Atlantic states (T2--747-749)

New York (T2--747)

New York (Manhattan Island, New York County) (T2--7471)

Grundnotation

coli-ana Workflow



Beispiel vc_daygram

```
700.90440747471 <liu_2_to_analyze; length: 15>
7----- Arts & recreation <dno_main>
70----- Arts <dno_div>
700----- The arts <dno_sec>
700.9----- Standard subdivisions of the arts #dno_span_cen# <dno_sub_span:700.1-700.9>
700.9----- History, geographic treatment, biography of the arts #dno_syn#
700.904----- Modern arts <RI_bui>
700.904----- Modern arts <dno_bui>
-0----- <Facet Indicator> <0>
--0----- Table 1. Standard Subdivisions <tabno:T1--0>
--0.9----- History, geographic treatment, biography <tabno:T1--09>
--0.9----- Regional treatment <RI:T1--09>
--0.904----- *20th century, 1900-1999 <tabno:T1--0904>
--0.904----- Historical periods #dno_span_cen# <tabno_span:T1--0901-T1--0905>
--0.904----- Twentieth century <RI:T1--0904>
--0.9044----- *1940-1949 <tabno:T1--09044>
--0.9044----- World War II, 1939-1945 <RI:T1--09044>
---.----07----- Museums, collections, exhibits; collecting objects <p9->tabno_span_1:T1--0901-T1--0905:07>
---.----074----- Museums, collections, exhibits <p9->tabno_span_1:T1--0901-T1--0905:074>
---.-----7----- North America <p20_5->tabno:T2--7>
---.-----7----- Modern world; extraterrestrial worlds #dno_span_cen# <p20_5->tabno_span:T2--4-T2--9>
---.-----7----- North America <p20_5->RI:T2--7>
---.-----74-- Northeastern United States (New England and Middle Atlantic states) <p20_5->tabno:T2--74>
---.-----74-- Specific states of United States #dno_span_cen# <p20_5->tabno_span:T2--74-T2--79>
---.-----74-- Northeastern States <p20_5->RI:T2--74>
---.-----747- New York <p20_5->tabno:T2--747>
---.-----747- Middle Atlantic states #dno_span_cen# <p20_5->tabno_span:T2--747-T2--749>
---.-----747- New York (State) <p20_5->RI:T2--747>
---.-----7471 New York <p20_5->tabno:T2--7471>
---.-----7471 New York Metropolitan Area <p20_5->RI:T2--7471>
```

Use Cases

- **Eintrag in den Katalogen** (z.B. K10plus) und bibliographischen Retrieval-Systemen
- **Erweiterung der Suchfunktionalität**
- **Aufbau von coli-ana Webservice**
- **Integration in das Mapping Tool Cocoda** (Mapping von semantischen Komponenten)
- **Qualitätskontrolle:** Bestimmung ungültiger DDC-Notationen.



coli-ana in K10plus Katalog

Titel: Optimierte Auftragsverfahren in der Spritzglastechnologie / Undine Fischer
Autorin/Autor: Fischer, Undine, 1968- 
Erschienen: Freiberg : Techn. Univ. Bergakad., 2009
Umfang: 89 S. : Ill., graph. Darst.
Sprache(n): Deutsch
Schriftenreihe: Freiburger Forschungshefte. Reihe A ; 897
ISBN: 978-3-86012-368-3
Sonstige Nummern: OCoLC: 436281776 →  WorldCat
OCoLC: 436281776 (aus SWB) →  WorldCat

RVK-Notation: ZM 6210  → Ähnliche Literatur

Sachgebiete: DNB- 666.444 (Grundnotation: 666.4) ; Not. anderer Haupttafeln 738.144




Molecular DDC Notation

Atomic DDC Notation

coli-ana Webservice

analyze Language: **Deutsch**, Norsk

Examples: 700.23, 700.90440747471, 666.444, 555.55

666.444 Glasieren   

666.444

6----- Technik, Medizin, angewandte Wissenschaften (600)

↳ 66----- Chemische Verfahrenstechnik (660)

↳ 666----- Keramiktechnologie und zugeordnete Technologien (666)

↳ 666.4-- Materialien, Ausstattung, Verfahren im Töpferhandwerk (666.4)

↳ **666.444 Glasieren--Töpferhandwerk (666.444)**

 ---.4- Techniken und Verfahren (738.14)

↳ ---.44 Glasieren (738.144)

PICA+: 045H/20 \$eDDC23ger\$a666.444\$c666.4\$d738.144\$Acoli-ana

Pica3: 5420 [DDC23ger]666.444-G--666.4-H--738.144\$Acoli-ana

API: **JSKOS** • **PICA/JSON** • **PICA Plain** • **Pica3**

deep link into K10plus catalog

atomic dno

analysis in the catalog record

analysis in machine-readable form (API)

Integration Mapping Tool Cocoda

★ **DDC Dewey Decimal Classification**

🔍 Type to search...

▼ **600 Technik**

- 660 Chemische Verfahrenstechnik
- 666 Keramiktechnologie und zugeordnete Technologien
- 666.4 Materialien, Ausstattung, Verfahren im Töpferhandwerk
- 666.44 Techniken und Verfahren
- ✖ **666.444 Glasieren** ★

Info **Labels** **Search** **Links** **coli-ana**

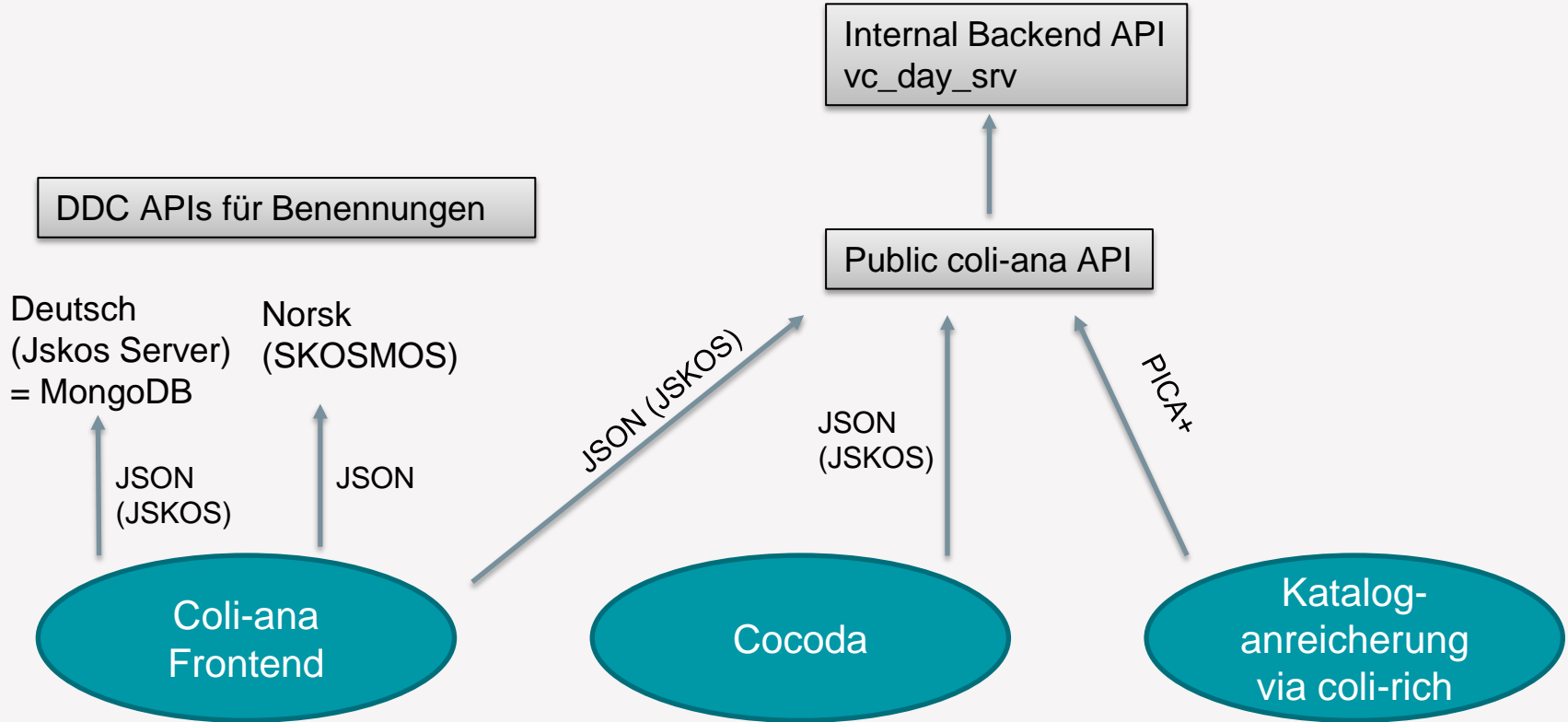
666.444

- 6 - - - - - Technik (600)
- ↳ 66 - - - - - Chemische Verfahrenstechnik (660)
- ↳ 666 - - - - - Keramiktechnologie und zugeordnete Technologien (666)
- ↳ 666.4 - - - **Materialien, Ausstattung, Verfahren im Töpferhandwerk (666.4)**
- 666.444 Glasieren (666.444)
- - - . -4 - Techniken und Verfahren (738.14)
- ↳ - - - . -44 **Glasieren (738.144)**

Go to the [coli-ana web interface](#) for more details and information.

Create and manage mappings between atomic DDC notations and entries of other vocabularies (Wikidata, LCSH, GND)

Technische Umsetzung



Anreicherung des K10plus Verbundkatalogs

1. **Database dump** or Abfrage in PICA+ format
2. **Extrahieren** von DDC aus den Titeldatensätzen

```
PICA+ 045F|045H  
MARC21 082
```

3. **Query Analyse** durch coli-ana API
4. **Erstellung** von PICA **patch** format

```
003@ $0600713679  
+ 045H/20 $eDDC23ger$a666.444$c666.4$d738.144$Acoli-ana
```

5. **Update** des Verbundkatalogs

coli-ana API

- Public endpoint at <https://coli-conc.gbv.de/coli-ana/app/analyze>
- Sources and documentation at <https://github.com/gbv/coli-ana>
- JSKOS result format (Komplette Zerlegung)
- PICA+ result format (auch alle Atom-Teile)

```
$ curl -s 'https://coli-conc.gbv.de/coli-ana/app/analyze?notation=666.444&format=picajson' \  
  | jq -r '[][6:-2][]' | grep [0-9]
```

666.4

738.144

Referenzen

<https://coli-conc.gbv.de/coli-ana/> coli-ana homepage

<https://coli-conc.gbv.de/publications/> publications of project colibri

Reiner (2016): Automatic Analysis of DDC Numbers based on MARC21

https://www.gbv.de/Verbundzentrale/Publikationen/publikationen-der-vzg-2016/pdf/reiner_160425_EDUG_Symposium.pdf

Reiner (2008): Automatic Analysis of Dewey Decimal Classification Notations

https://doi.org/10.1007/978-3-540-78246-9_82

https://www.gbv.de/Verbundzentrale/Publikationen/2008/2008/pdf/pdf_3936.pdf

Anprechpartner

Dr. Ulrike Reiner

ulrike.reiner@gbv.de

Dr. Jakob Voß

jakob.voss@gbv.de

Uma Balakrishnan

uma.balakrishnan@gbv.de

Stefan Peters

stefan.peters@gbv.de





**Vielen Dank!
Fragen?**