coli-ana

Automatic analysis of the Dewey Decimal Classification
A service of the Verbundzentrale GBV

Uma Balakrishnan / Stefan Peters / Jakob Voß
Verbundzentrale des GBV (VZG)

SWIB conference 2021-12-01

Agenda

- Colibri
- Dewey Decimal Classification (DDC)
- coli-ana: automatic analysis of the DDC numbers
- Challenges
- Workflow
- Use cases

Initiated by Dr. Ulrike Reiner (2003)

Colibri Project

Funded by the VZG

Context Generator & Linguistic Tools for Bibliographic Retrieval Interface

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?

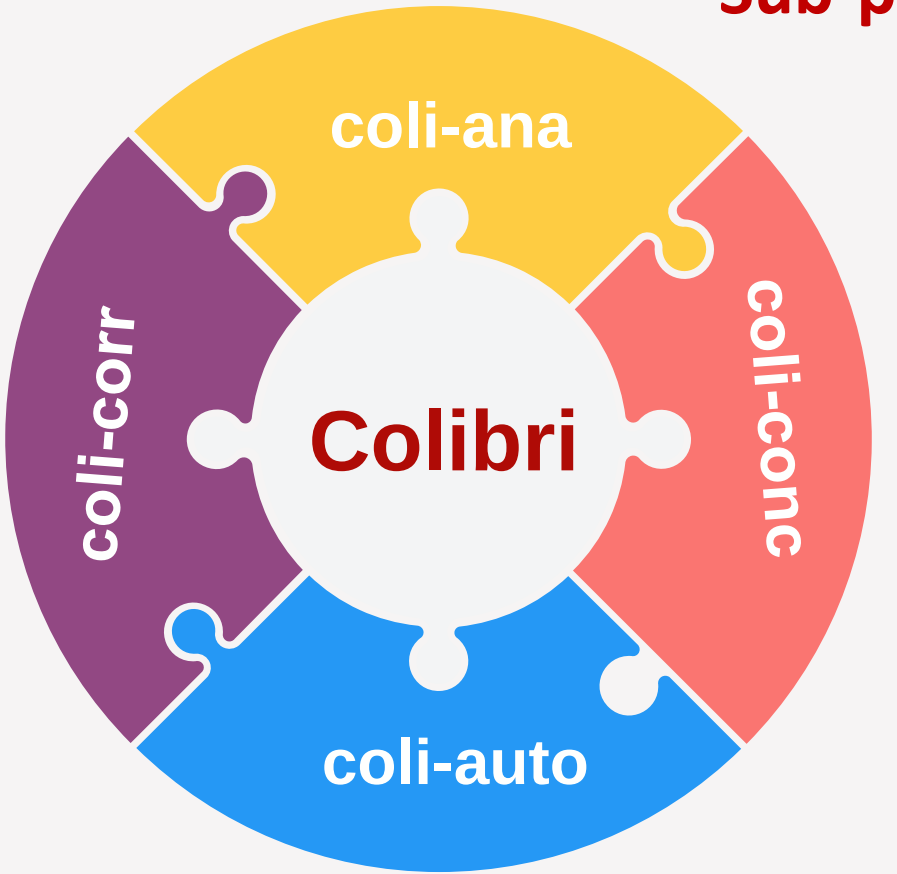
Q3. …**improve automatic classification & retrieval** by means of atomic DDC notations?

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

Sub-projects



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

DDC

- **Actively in use** over a century
- **Large user community** worldwide
- VZG member of the **Dewey Consortium** in 2000
- **Strong representation** in Europe: EDUG User group
- **Dynamic system**
- **Rich system, precisely structured notations**
- **Huge influx** of Dewey numbers into the K10lus Catalog from external data
- **At least 1 Mio. unique DDC** built numbers in K10plus Catalog

DDC System and numbers

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

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

100 Philosophy & psychology
100 **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

Complexity of the DDC numbers

331.89282922520971274309051 1

700.9044074747

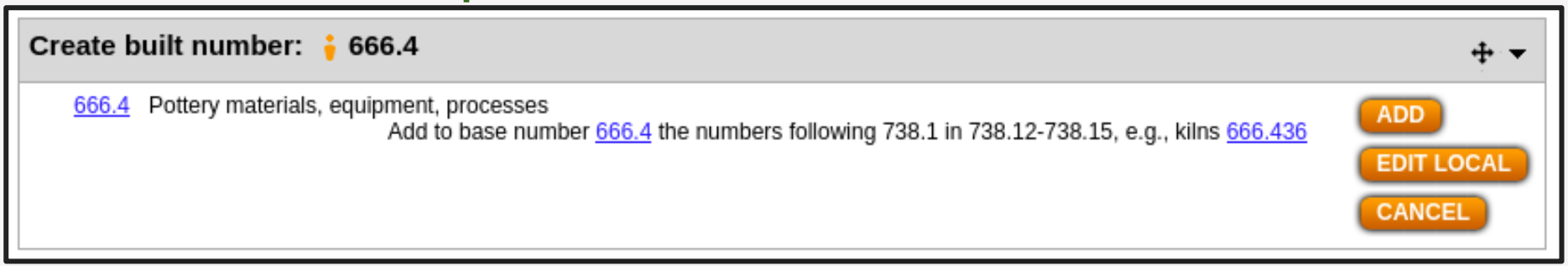
754.09109033

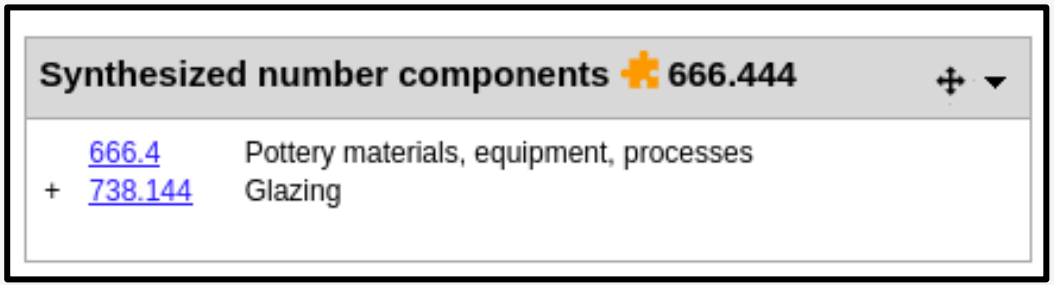
700.23

700

DDC number building

Finely structured and precise numbers can be composed from the multiple parts of the DDC **based on complex rules**



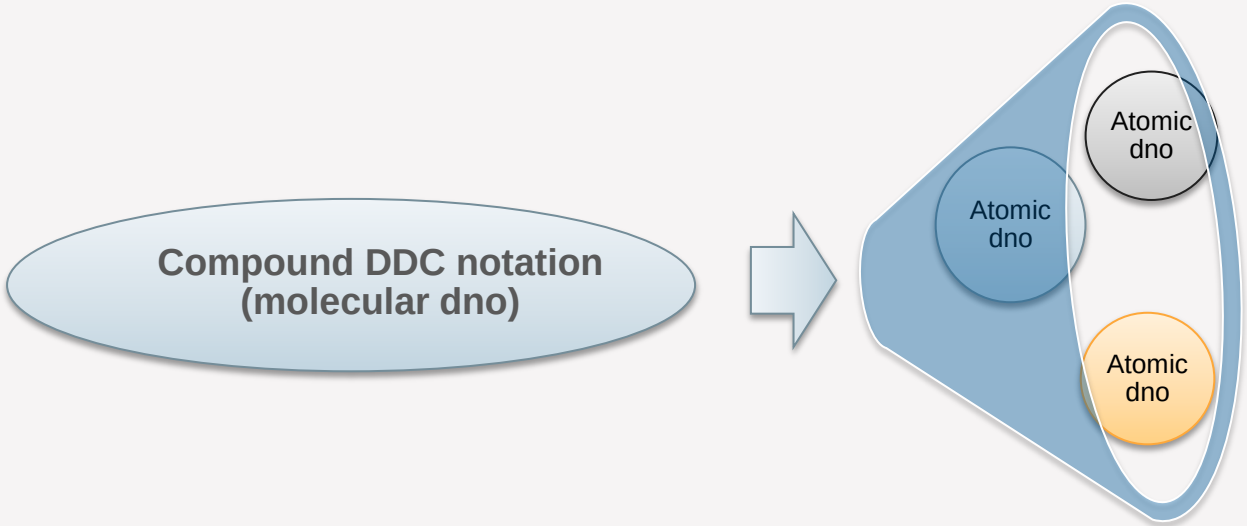


Pottery glazing technology

666.4 + 738.1**44** ⇒ **666.444**

coli-ana

A tool for automatic analysis of synthesized DDC notations



Objectives

Improve Retrieval

- Support of search terms (atomic dnos)
  Example: T1--09044 (DDf *1940-1949) returns all titles that are in any way related to the 1940s
- Extension to full text search through captions (all captions contained in a sythesized dno)
- Assessment and ranking of similar publications

Improve the presentation in the catalogue

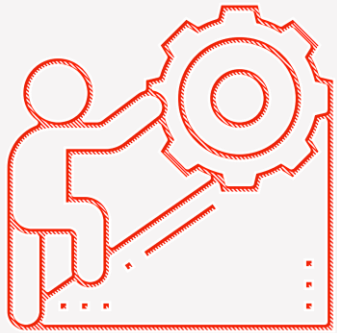
The cryptic DDC notation to the titles can be enriched with captions and presented similar to the Regensburg Verbund Classification in the K10plus Catalog

Analyse & enhance subject indexing

The components of an analysis can also be mapped individually with elements of other systems. These mappings can then be used for enrichment and to examine subject indexing.
Example: DDC 700 "Arts" and T1--0901-0905:074 "Museums, collections, exhibitions" => BK 20.13 "Art exhibition".

Challenges in automatic analysis

- **Extensive System** with over **51.700** classes
- **Complex number building** system
  - Main schedule, six tables and other auxillary tables
  - Standard subdivisions, notes
  - **Over 9.987 instructions** (e.g. add note, authorisation, discontinuation, include note, class here, class elsewhere, revision,…)
    - Bundled into 60 rule parts
- **Possibility to build** fine and accurate DDC numbers that can get very lengthy

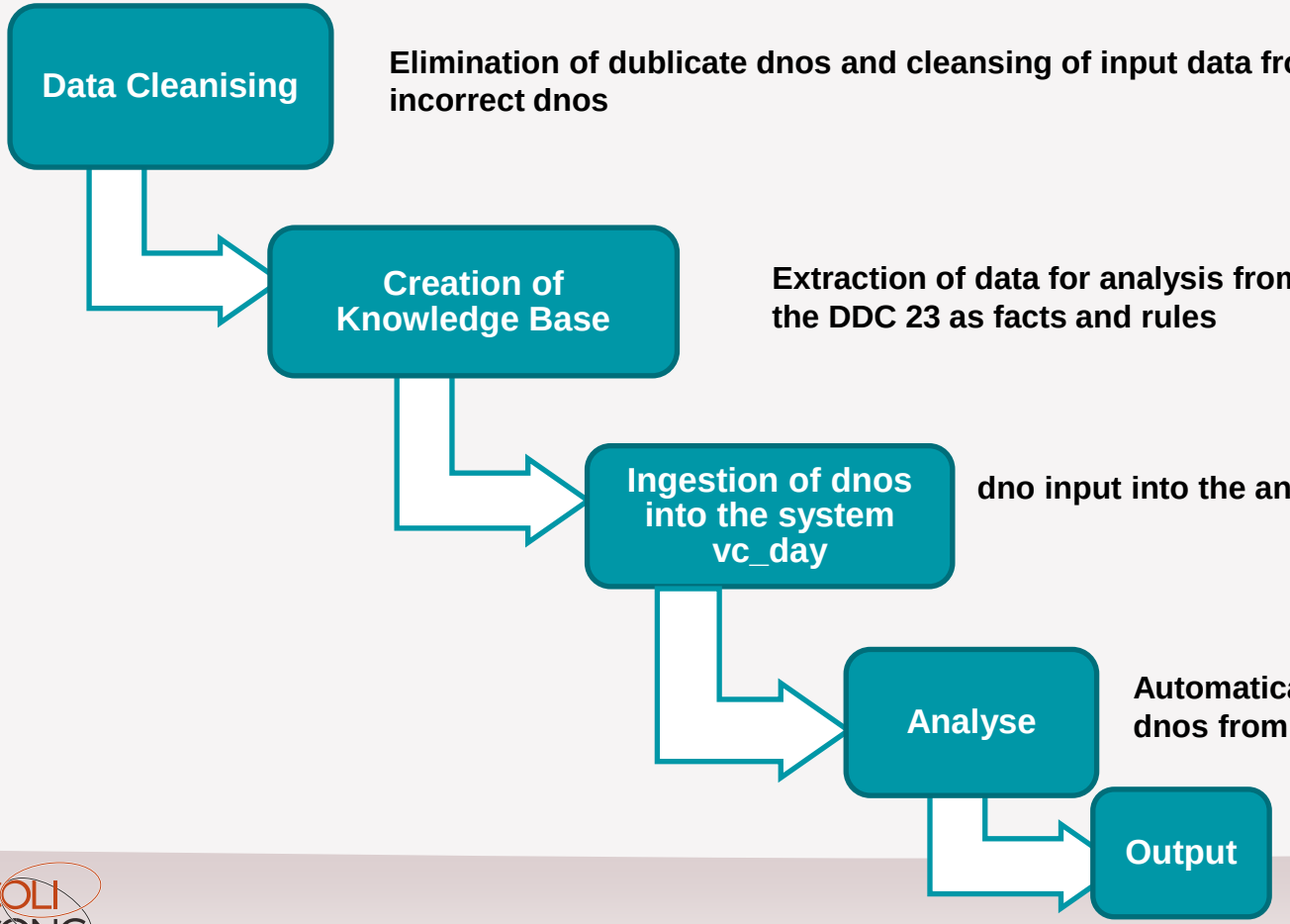


A complete analysis of a rich DDC number

700.90440747471	
7----------	Arts & recreation (700)
70---------	Arts (700)
700--------	The arts (700)
700.9------	Standard subdivisions of the arts (700.1-700.9)
700.9------	History, geographic treatment, biography of the arts (700.9)
700.904-----	Arts--20th century,... (700.904)
-0---------	facet indicator (0)
--0--------	Table 1. Standard Subdivisions (T1--0)
--0.9------	History, geographic treatment, biography (T1--09)
--0.904-----	Historical periods (T1--0901-0905)
--0.904-----	*20th century, 1900-1999 (T1--0904)
--0.9044----	*1940–1949 (T1--09044)
---.----07---	Museums, collections, exhibits; collecting objects (T1--0901-0905:07)
---.----074--	Museums, collections, exhibits (T1--0901-0905:074)
---.-------7-	Modern world; extraterrestrial worlds (T2--4-9)
---.-------7-	North America (T2--7)
---.-------74	Specific states of United States (T2--74-79)
---.-------74	Northeastern United States (New England and Middle Atlantic states) (T2--74)
---.-------747	Middle Atlantic states (T2--747-749)
---.-------747	New York (T2--747)
---.-------7471	New York (Manhattan Island, New York County) (T2--7471)

Base number

coli-ana Workflow

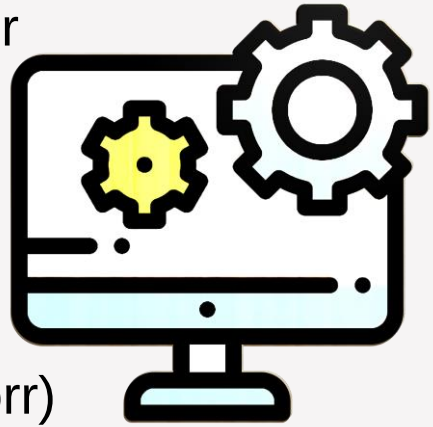


Example vc_daygram

```
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

- **Entry** into the Catalogues (for e.g. K10plus) and other bibliographic retrieval systems
- **Extend** search functionality
- **Analyse and Re-use** (coli-ana webservice)
- **Map** semantic components (Mapping Tool Cocoda)
- **Quality** Control: detect invalid DDC notations (coli-corr)

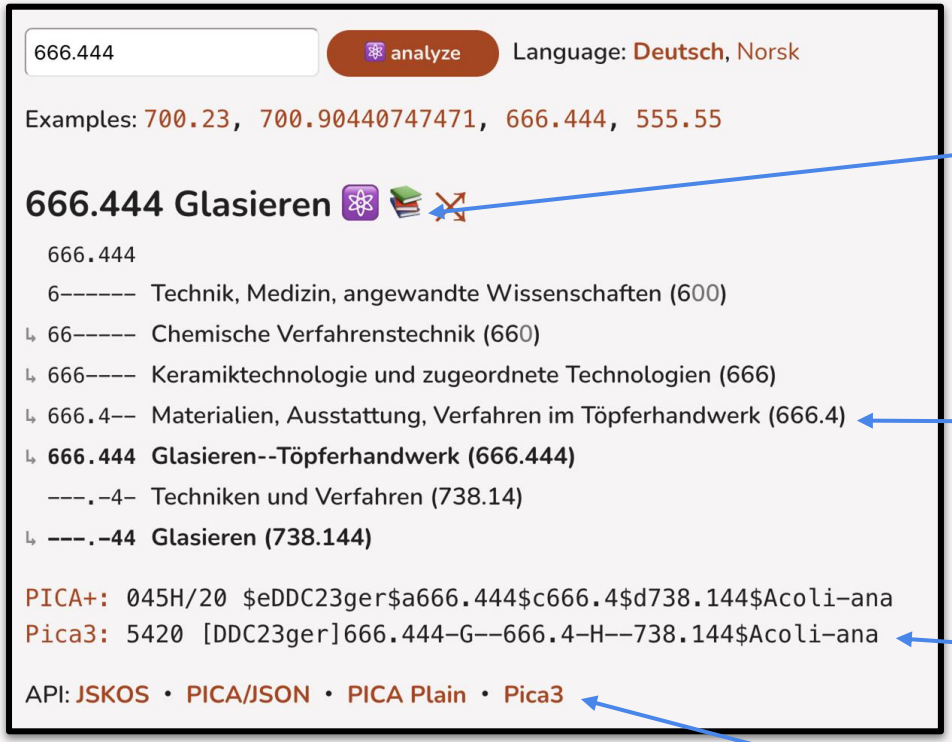


coli-ana in K10plus catalog

molecular DDC notation

atomic DDC notations

coli-ana webservice



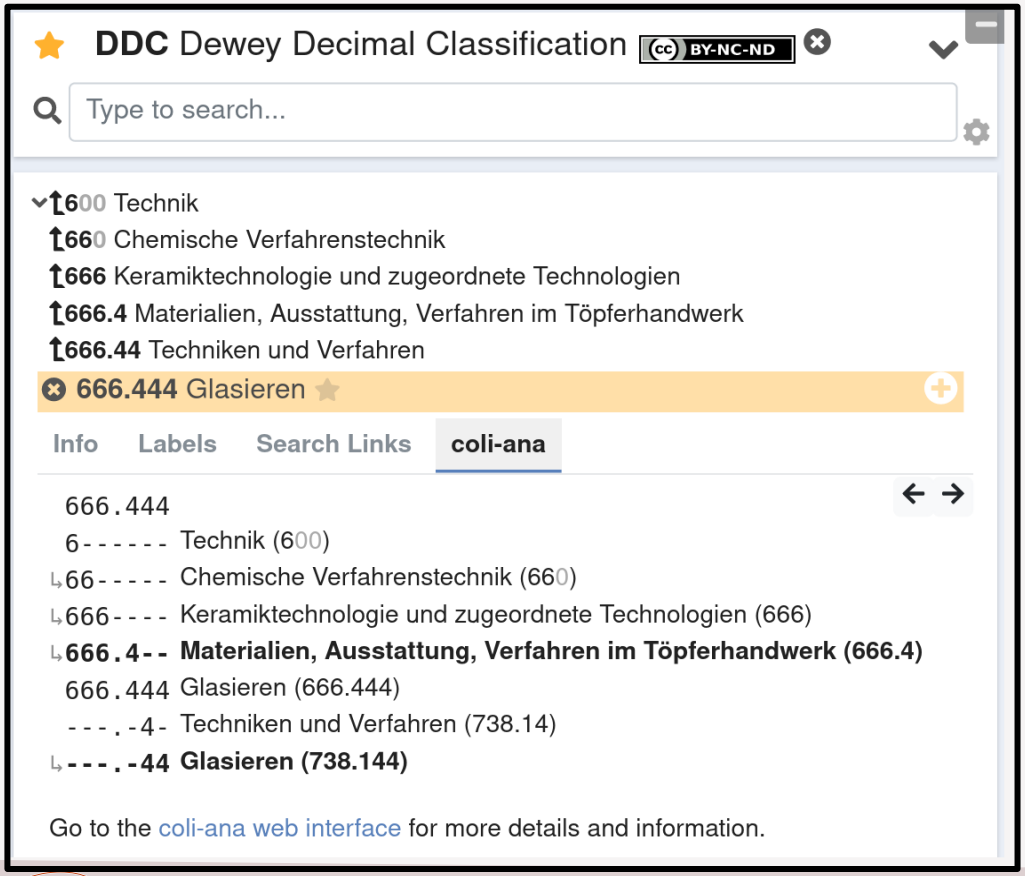
deep link into K10plus catalog

atomic dno

analysis in the catalog record

analysis in machine-readable form (API)

Cocoda mapping tool integration



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

References

Visit us at SWIB booth for discussion!

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

Contact

Dr. Ulrike Reiner
ulrike.reiner@gbv.de

Uma Balakrishan
uma.balakrishnan@gbv.de

Dr. Jakob Voß
jakob.voss@gbv.de

Stefan Peters
stefan.peters@gbv.de

Thank You!
Questions?