

# Lessons learned while building the vocabulary mapping tool Cocoda

ELAG Berlin

Jakob Voß   Stefan Peters   Uma Balakrishnan

2019-05-08

# Introducing coli-conc

# Colibri, the mother of all

**C**ontext generation and **l**inguistic tools for **b**ibliographic retrieval interfaces, conducted by our colleague Ulrike Reiner since 2002

- ▶ automatic DDC classification (**coli-auto**)
- ▶ automatic checking of DDC number correctness (**coli-corr**)
- ▶ automatic analysis and synthesis of DDC numbers (**coli-ana**)
- ▶ semi-automatic creation of DDC concordance (**coli-conc**)

⇒ Domain expert and previous works to build on

# coli-ana, an example

700.90440747471



Info

Search Links

coli-ana

700.90440747471	
7	Arts & recreation
70	Arts
700	The arts
700.904	Modern arts
T1--0904	20th century, 1900-1999
T1--09044	1940-1949
T1--074	Museums, collections, exhibits
T2--7	North America
T2--74	Northeastern United States (New England and Middle Atlantic states)
T2--747-749	Middle Atlantic states
T2--747	New York (State)
T2--7471	New York Metropolitan Area



## coli-conc, an example

The screenshot shows a Cocoda interface window with a title bar containing a gear icon and a minus sign. The main content area is divided into two columns: "DDC" on the left and "Wikidata" on the right. In the DDC column, the text "700.904 Künste—20. Jahrhundert" is displayed with a close button (⊗) to its right. In the Wikidata column, the text "Q4630639 20th-century art" is displayed with a close button (⊗) to its right. An approximation symbol (≈) is positioned between the two terms. At the bottom of the window, there is a toolbar with icons for a speech bubble, a double-headed arrow, a document, a trash can, and a prohibition sign. The name "Jakob Voß" and a code icon (</>) are visible in the bottom right corner.

## coli-conc, the idea

- ▶ a **mapping-tool** to avoid aching hands
- ▶ facilitate creation and management of **concordances** between **knowledge organization systems (KOS)**

## coli-conc, the idea

- ▶ a **mapping-tool** to avoid aching hands
- ▶ facilitate creation and management of **concordances** between **knowledge organization systems (KOS)**
- ▶ in particular library classifications
  - ▶ Dewey Decimal Classification (DDC)
  - ▶ Regensburger Verbundklassifikation (RVK)
  - ▶ Basisklassifikation (BK)
  - ▶ *several local classification schemes*
  - ▶ ...

## coli-conc, to-do list

1. Collect KOS metadata  
⇒ BARTOC registry



## coli-conc, to-do list

1. Collect KOS metadata  
⇒ BARTOC registry
2. Collect and publish existing mappings
  - ▶ DDC/RVK/GND/BK/STW/LCSH/lxTheo  
384.491 mappings
  - ▶ Wikidata  
3.607.683 mappings (as of 6/2018)
  - ▶ additional mappings not converted yet

## coli-conc, to-do list

1. Collect KOS metadata  
⇒ BARTOC registry
2. Collect and publish existing mappings
  - ▶ DDC/RVK/GND/BK/STW/LCSH/IxTheo  
384.491 mappings
  - ▶ Wikidata  
3.607.683 mappings (as of 6/2018)
  - ▶ additional mappings not converted yet
3. Create mapping tool  
⇒ Cocoda!

# Cocoda 2014: first prototype with AngularJS

Cocoda *prototype under construction* Log in

Source Scheme: DDC -

Search Options ▾

Search by: **Term** Notation

**612.112** Leukozyten (Weiße Blutkörperchen)

- ┆ Blut
- ┆ Biochemie
- ┆ Biophysik
- ┆ Anzahl und Auszählung

Map → Look up database all ▾

Active Mapping

612.112 ⓘ Ⓞ → No target concepts selected!

Target Scheme: RVK -

Search Options ▾

Search by: **Term** Notation

**WW 8840 - WW 8879**

Add + Replace all →

Mapping Candidates ✕

Catalog Occurrences ▾

Used notation: **612.112**  
Used database: **GVK/SWB**  
Results (total) for **612.112**: 42  
Corresponding notations in **RVK**:

Notation	Hits	% of total
<b>WW 8840</b> ⓘ +	22	52.4 %
<b>YC 2500 - YC 2599</b> ⓘ +	11	26.2 %
<b>WF 9895</b> ⓘ +	8	19.0 %
<b>XG 6700 - XG 6728</b> ⓘ +	1	2.4 %

Top Concepts ▾

- A Allgemeines ⓘ
- B Theologie und Religionswissenschaften ⓘ
- CA - CK Philosophie ⓘ
- CL - CZ Psychologie ⓘ
- D Pädagogik ⓘ
- E Allgemeine und vergleichende Sprach- und Literaturwissenschaft, Indogermanistik, Außereuropäische Sprachen und Literaturen ⓘ
- F Klassische Philologie, Byzantinistik, Mittellateinische und Neuarische

# Cocoda 2015-2016

- ▶ Write project grant and receive additional funding
- ▶ Continue with specification (esp. JSKOS data format)
- ▶ Create a new implementation

# JSKOS data format for Knowledge Organization Systems

- ▶ JSON-LD for SKOS (prefLabel, broader, narrower...)
- ▶ Additional properties from other ontologies (url, next, previous, startDate, endDate...)
- ▶ Additional classes for mappings, concordances, registries...

⇒ <https://gbv.github.io/jskos/>

# Getting data into JSKOS format

1. CSV, MARCXML, SKOS, ... → Cleanup
2. CSV, MARCXML, SKOS → JSKOS

Using at least three different tools:

- ▶ skos2jskos (Perl)
- ▶ jskos-convert (NodeJS)
- ▶ mc2skos (Python)
- ▶ scripts for minor JSON adjustments (jq)

- ▶ Write project grant and receive additional funding
- ▶ Continue with specification (especially JSKOS data format)
- ▶ Create a new implementation

- ▶ Write project grant and receive additional funding
- ▶ Continue with specification (especially JSKOS data format)
- ▶ Create a new implementation
  - ▶ as monolithical Java application



- ▶ Write project grant and receive additional funding
- ▶ Continue with specification (especially JSKOS data format)
- ▶ Create a new implementation
  - ▶ as monolithical Java application
  - ▶ and throw it away afterwards

# coli-conc 2018: start new from scratch (Node & Vue)

The screenshot displays the Cocoda Prototype interface for comparing two entities: 'DQ Lehrpläne' and '81.62 Curriculum'. The interface is divided into several panels:

- Left Panel (RVK Regensburger Verbundklassifikation):** A tree view showing categories like 'Pädagogik', 'Lehrpläne', and 'Berufsbildung'. 'DQ Lehrpläne' is selected.
- Top Panel:** Shows the entity names 'DQ Lehrpläne' and '81.62 Curriculum' with a comparison icon.
- Comparison Table:** A table with columns 'from', 'to', and 'creator'. It lists various relationships between 'DQ' and 'BK' (Basisklassifikation) entities, including 'Local (1 / 2)', 'Registry', 'Wikidata', and 'Occurrences'.
- Right Panel (Basisklassifikation):** A tree view showing the hierarchy of '81.62 Curriculum', including '7-8 Sozialwissenschaften' and '1.100 Unterrichtsprozeß: Allgemeines'.

from	to	creator
RVK DQ	BK 81.62	You
RVK GU 10800	BK 81.62	ULB Tirol
DDC 375	BK 81.62	VZG
DDC 375	RVK DQ	VZG
?	BK 81.62	GVK - G... 2561
RVK DQ	BK	GVK - G... 201
RVK DQ	DDC 370	GVK - G... 30
RVK DQ	BK 81.62	GVK - G... 22
RVK DQ	RVK DQ 3018	GVK - G... 22
RVK DQ	RVK DQ 3406	GVK - G... 21
RVK DQ	RVK DQ 3414	GVK - G... 20
RVK DQ	RVK DQ 3416	GVK - G... 20

# coli-conc 2019: curent layout

Cocoda Mapping Tool (dev)

Help Feedback You

**RVK Regensburger**

Verbundklassifikation [6] **multimedia**

Lehrpläne

D Pädagogik

**DQ Lehrpläne**

Info Search Links

<http://rvk.uni-regensburg.de/ht/DQ>

19863:

Label (de): Lehrpläne

Created: 5. Jul 2012

Modified: 19. März 2019

[ DQ 1000 - DQ 4820 Deutsche Lehrpläne

[ DQ 4900 Lehrpläne Italien

[ DQ 5000 - DQ 5009 Ausländische Lehrpläne

UL Berufsberatung, Berufsberatung Österreich

DM Universität, Hochschule, Forschung und Wissenschaft

DN Lehrer und Lehrerbildung

DO Spezialfragen des gesamten Schulsystems

DP Didaktik und Methodik des Unterrichts

**DQ Lehrpläne**

DR Erziehungsrrecht

DS Sozialpädagogik, Sozialarbeit

DT Heilpädagogik, Sonderpädagogik

**Mapping Editor**

RVK ≈ BK

DQ Lehrpläne ≈ 81.62 Curriculum

Mapping Browser

Available Mappings Available Automatic Mappings

	From	To	Creator
L	RVK DQ Lehrpläne	= BK 81.62 Curriculum	You
C	RVK GU 10800 Curriculumforschung, Empirische Fremdsprachenunterrichtsforcl Lehrerbildung	BK 81.62 Curriculum	ULB Tirol
C	DDC 375 Curricula	- 81.62 Curriculum	VZG
C	- 375 Curricula	RVK DQ Lehrpläne	VZG
W	WD Q207137 Curriculum	BK 81.62 Curriculum	
CO	— —	BK 81.62 Curriculum	GVK - ... 2571
CO	RVK DQ Lehrpläne	— —	GVK - ... 203
CO	- DQ Lehrpläne	DDC 370 Bildung und Erziehung	GVK - ... 31

**BK Basisklassifikation**

Curriculum

7-8 Sozialwissenschaften

81.60 Unterrichtsprozess: Allgemeines

**81.62 Curriculum**

Editorial Register Entries Info

Search Links

<http://uni.gbv.de/terminology/bk/81.62>

Label (de): Curriculum

Modified: 29. Nov. 2018

Schulpsychologie: Allgemeines

81.60 Unterrichtsprozess: Allgemeines

81.61 Didaktik, Hochschuldidaktik

**81.62 Curriculum**

81.64 Unterrichtsprozess: Sonstiges

81.65 Lehrmittel, Lernmittel: Allgemeines

81.70 Bildungssysteme, Bildungsinstitutionen: Allgemeines

81.99 Bildungswesen: Sonstiges

83 Volkswirtschaft

85 Betriebswirtschaft

86 Recht

# Live Demo

`https://coli-conc.gbv.de/cocoda/app/`

# Infrastructure

- ▶ jskos-server & DANTE terminology registry (JSKOS-API)
- ▶ mapping suggestions (OpenRefine Reconciliation API)
- ▶ login-server (OAuth)

# Lessons learned

# People stick to spreadsheets

- ▶ limited
  - ▶ 1-to-n mappings
  - ▶ repeatable fields
  - ▶ leading zeroes
  - ▶ ...
- ▶ at least better than MS Word
- ▶ you think CSV, they think Excel

# People stick to spreadsheets

- ▶ limited
  - ▶ 1-to-n mappings
  - ▶ repeatable fields
  - ▶ leading zeroes
  - ▶ ...
- ▶ at least better than MS Word
- ▶ you think CSV, they think Excel

⇒ ***spreadsheets are not perfect but actually useful***



# Software development is communication

- ▶ listen
  - ▶ what is actually wanted?
  - ▶ how do the involved parties work?
- ▶ understand
  - ▶ face2face meetings to find a common language
  - ▶ most problems are communication problems
- ▶ explain
  - ▶ pros and cons of technical decisions
  - ▶ basic such as URIs and Open Data
  - ▶ not all features are implemented first

# Software development is communication

- ▶ listen
  - ▶ what is actually wanted?
  - ▶ how do the involved parties work?
- ▶ understand
  - ▶ face2face meetings to find a common language
  - ▶ most problems are communication problems
- ▶ explain
  - ▶ pros and cons of technical decisions
  - ▶ basic such as URIs and Open Data
  - ▶ not all features are implemented first

⇒ ***Communicate!***

## No schema, no data quality

- ▶ Notations and identifiers must match regular expressions
- ▶ JSON Schema helped to find inconsistencies in JSON data
- ▶ Additional constraints not expressible in JSON Schema
- ▶ empty strings, empty arrays, null values...

## No schema, no data quality

- ▶ Notations and identifiers must match regular expressions
- ▶ JSON Schema helped to find inconsistencies in JSON data
- ▶ Additional constraints not expressible in JSON Schema
- ▶ empty strings, empty arrays, null values...

⇒ ***Never trust any data you haven't validated!***

# Holy decoupling

- ▶ service oriented architecture (SOA)
- ▶ APIs and data formats matter most
- ▶ Things will break anyway

# Holy decoupling

- ▶ service oriented architecture (SOA)
- ▶ APIs and data formats matter most
- ▶ Things will break anyway

Easy to replace parts of the infrastructure

- ▶ RVK API → own DB → DANTE → jskos-server
- ▶ PHP → Perl → JavaScript

# Holy decoupling

- ▶ service oriented architecture (SOA)
- ▶ APIs and data formats matter most
- ▶ Things will break anyway

Easy to replace parts of the infrastructure

- ▶ RVK API → own DB → DANTE → jskos-server
- ▶ PHP → Perl → JavaScript

⇒ ***split your application into decoupled services!***

## Look out for beneficial beta-users

- ▶ users prefer a mature product instead buggy prototype
- ▶ real use-cases and outcome instead of click-around testing
- ▶ use-cases will drive development in different directions



## Look out for beneficial beta-users

- ▶ users prefer a mature product instead buggy prototype
- ▶ real use-cases and outcome instead of click-around testing
- ▶ use-cases will drive development in different directions

⇒ ***Agile development requires some agile users!***

# Patience and luck

- ▶ Colibri started in 2002
- ▶ coli-conc started in 2013
- ▶ Current development cycle started in 2018

# Patience and luck

- ▶ Colibri started in 2002
- ▶ coli-conc started in 2013
- ▶ Current development cycle started in 2018

⇒ ***Good luck!***

# Summary

## We build...

- ▶ JSKOS data format to unify KOS & concordances data
- ▶ several web services to process and share mapping data
- ▶ a web application based on the format and services

## We learned to...

- ▶ stop worrying about spreadsheets
- ▶ listen, understand, and explain
- ▶ insist on schemas for data quality
- ▶ decouple and rewrite services
- ▶ work together with beta-user
- ▶ be more patient and use lucky chances

## Feedback is welcome!

<https://coli-conc.gbv.de/> the project

<https://coli-conc.gbv.de/cocoda/> the mapping-tool

<https://github.com/gbv/cocoda/> the source code

<https://gbv.github.io/cocoda/> the documentation

<https://github.com/gbv/cocoda/issues> the issue tracker