European Conference on Data Analysis
Gfkl Workshop
3. September 2015, University of Essex

Cocoda “Colibri Concordance Database“ –
A mapping tool for library classification schemes

U. Balakrishnan, Verbundzentrale des GBV
Overview

- Background of the Project coli-conc
- Methods of Mapping
- Course Correction
- Introduction to the Software
  - Demands on the Tool
  - Web Layout
  - Software Concept
## Classification systems in German speaking regions

<table>
<thead>
<tr>
<th>Universal Classification Systems</th>
<th>No. of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>UDC (Universal Decimal Classification)</td>
<td>ca. 65.000 classes (English version)</td>
</tr>
<tr>
<td>DDC (Dewey Decimal Classification)</td>
<td>over 44,000 classes with 10 main classes</td>
</tr>
<tr>
<td>RVK (Regensburg Classification)</td>
<td>850,000 classes with 33 main classes</td>
</tr>
<tr>
<td>BC (Basic Classification)</td>
<td>2100 classes with 89 main classes</td>
</tr>
<tr>
<td>LCC (Library of Congress Classification)</td>
<td>21 main classes</td>
</tr>
</tbody>
</table>

### Subject classification

<table>
<thead>
<tr>
<th>No. of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDC-Sachgruppen der DNB</td>
</tr>
<tr>
<td>MSC (Mathematics Subject Classification)</td>
</tr>
<tr>
<td>PACS (Physics and Astronomy Classification Scheme)</td>
</tr>
<tr>
<td>FKDigBib (Subject classification for digital library)</td>
</tr>
<tr>
<td>KfM (Classification for music library)</td>
</tr>
</tbody>
</table>

### Subject Classification at the Universities

<table>
<thead>
<tr>
<th>No. of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>TUM-classification (Science and technology classification of the TU Munich)</td>
</tr>
<tr>
<td>Subject classification of the University library Duesseldorf</td>
</tr>
<tr>
<td>Bremer classification of the State and University library Bremen</td>
</tr>
<tr>
<td>GOK (Goettingen Online Classification)</td>
</tr>
<tr>
<td>Standard-Thesaurus Wirtschaft von der ZWB</td>
</tr>
<tr>
<td>Subject classification University library Trier</td>
</tr>
<tr>
<td>Technical University Dortmund</td>
</tr>
<tr>
<td>University library Paderborn</td>
</tr>
<tr>
<td>University library Marburg</td>
</tr>
<tr>
<td>University library Bonn</td>
</tr>
<tr>
<td>University library Heidelberg</td>
</tr>
<tr>
<td>Subject classification and nomenclature of individual languages Library of the Institute of General Linguistics at the Uni Münster</td>
</tr>
</tbody>
</table>

### Subject Classification at the public libraries

<table>
<thead>
<tr>
<th>No. of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEB (Scheme for protestant libraries)</td>
</tr>
<tr>
<td>SKB-E (Scheme for catholic public libraries)</td>
</tr>
<tr>
<td>KfKJ (Scheme for children and youth libraries)</td>
</tr>
<tr>
<td>ASB (General classification for public libraries)</td>
</tr>
<tr>
<td>ÖSÖB (Austrian classification for public libraries)</td>
</tr>
<tr>
<td>SfB (Classification for libraries)</td>
</tr>
<tr>
<td>KAB (Classification for general libraries)</td>
</tr>
<tr>
<td>SSD (Classification of the city library Duisburg)</td>
</tr>
<tr>
<td>ESSB (Single classification for South Tyrolean)</td>
</tr>
</tbody>
</table>
Primary Source and Target Schemes: DDC and RVK

Why RVK?
- Wide-spread in Germany
- Local needs are better covered
- Legacy data transfer
- DDC is subject to licence

DDC
- 850,000 classes
- 33 main classes
- Granularity varies in different subject fields
- Synthesized notations are prebuilt and integrated into the online system

RVK
- ca. 40 Mio. Title data records (2013)
- 19.8% DDC, 3.7% RVK
- 14.4% DDC, 24.3% RVK

SWB
- ca. 17 Mio. Title data records (2013)
- 17.7% DDC, 29.2% RVK

BVB
- ca. 20 Mio. Title data records (2013)
- 17.7% DDC, 29.2% RVK

VZG
www.gbv.de
Survey

- Current status of DDC-X-concordance
- Field of application and the reasons for the use of the DDC
- Methods & Problems in building a DDC-X concordance
- Interest in a DDC - RVK concordance

### Existing Mapping works

<table>
<thead>
<tr>
<th>Concordance</th>
<th>Subject area</th>
<th>Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>DDC – BK</td>
<td>Chemistry</td>
<td>TUB TUHH</td>
</tr>
<tr>
<td></td>
<td>Politics</td>
<td>SUB Hamburg</td>
</tr>
<tr>
<td></td>
<td>The thousand classes of the third summary</td>
<td>VZG</td>
</tr>
<tr>
<td>DDC – EWB</td>
<td>41 EWB-Fachgruppen</td>
<td>VZG</td>
</tr>
<tr>
<td>DDC – RVK</td>
<td>Library- and Information science</td>
<td>HdM Stuttgart</td>
</tr>
<tr>
<td></td>
<td>Social science</td>
<td>UB Greifswald</td>
</tr>
<tr>
<td></td>
<td>Medicine &amp; Health, Law, the thousand classes of the third summary level</td>
<td>VZG</td>
</tr>
<tr>
<td>RVK – DDC</td>
<td>Biology, Chemistry, Geology, Paleontology, Phisics, Mathematics</td>
<td>GESIS</td>
</tr>
<tr>
<td></td>
<td>Psychology</td>
<td>SLUB Dresden</td>
</tr>
<tr>
<td>RVK – BK</td>
<td>German literature, Politics, Law</td>
<td>UB Wien</td>
</tr>
<tr>
<td>RVK – MSC</td>
<td>Mathematics</td>
<td>UB Regensburg</td>
</tr>
<tr>
<td>RVK – PACS</td>
<td>Physics</td>
<td>UB Regensburg</td>
</tr>
<tr>
<td>SWD – DDC</td>
<td>Library- and Information science</td>
<td>DNB</td>
</tr>
<tr>
<td>SWD-RVK</td>
<td>Library- and Information science</td>
<td>HdM Stuttgart</td>
</tr>
<tr>
<td>RVK-BK-MSC-PACS</td>
<td>Mathematics, Physics</td>
<td>ULB Tirol</td>
</tr>
<tr>
<td>DDC-MSC-BKL</td>
<td>Mathematics</td>
<td>TIB Hannover</td>
</tr>
</tbody>
</table>
Mapping Methods

**Work done so far:**
- Complete concordance
  - DDC - EZB
  - DDC - BK for the thousand classes of the third summary of the DDC
  - DDC - RVK for the thousand classes of the third summary of the DDC
- Partial Concordance
  - DDC - RVK for the DDC subject area „Law“ (ca. 14% of the current DDC-classes)
- DDC - RVK for the DDC subject area „Philosophy“

**Statistical Inference - Title data records**
- Catalogues and databases e.g. GVK, SWB

**Classification system based search**
- Term definition
- Synonym search

**Decision Making**
- Experience
- Expertise
Facilitate exchange and use of concordances and KOS
  - Collection of existing mappings and KOS
  - Provision of the above

Enhance the speed of building concordances between library KOS and ease their management
  - Develop a mapping tool
  - Make the concordances and KOS easily accessible
  - Draft algorithms for automatic generation of mapping candidates

Improve the Quality of the concordances
  - Develop and implement measures for quality control
  - Involve and expand the user groups
Demands on the tool

Integration of Data from different sources

Allow validation and storage of data

Presentation of Data and mapping candidates on a single screen

Multi-user web based open source tool

Easy access to and exchange of information
Serve as collaboration platform

Clear overview of the context of the selected term through display of
- the hierarchical structure of the classes
- scope notes
- Register Index Entries
- linked vocabularies and synonym suggestions

Mapping suggestions through
- evaluation of the co-occurrences of assigned notations/terms in the title data records
- automatic generation of mappings
- integration of the concordance database
- inclusion of the results of a manual mapping
Cocoda – Software Concept

Library knowledge organisation systems

Conversion
DDC-MARCXML in JSON format

DDC-Database
Couch DB in JSKOS format

RVK-SRU API
 cumulative search

GND-API lbbib.org

WIKIDATA-API

WIKIDATA- API wrapper
JSKOS API
not implemented yet.

SWB-Catalog SRU-API

GVK-Catalog SRU-API

Format requirements
- Compatible with SKOS
- Represent both KOS and its mappings
- Simple and easy to use in web application

JSKOS Format
- An application of JSON-LD doesn’t contain the full complexity of RDF
- Extended statements on mappings, e.g. creator, mapping methods
- Existence statements to express negation or completion of a KOS, e.g. no narrow concept exists or a broader terms exist

OpenSearch
suggestion module

KOS-
representation module

Catalog
occurrences module
not implemented yet.

Mapping
recommendation module
not implemented yet.

User
management module
not implemented yet.

Quality
assessment module
not implemented yet.

Concordance
Database
Couch DB in JSKOS format

Web interface

www.gbv.de
Thank You!

Vectors slide no.7: © Vallepu – fotolia.com [https://de.fotolia.com/]
Vectors slide no.5: © NLshop—fotolia.com [https://de.fotolia.com/]
Vectors slide no.2,3,4,6,8,14: designed by Freepik.com [http://www.freepik.com]
Thanks to Jana Agne for creating the table at the slide no.3