

Automatic creation of mappings between classification systems for bibliographic data

Prof. Magnus Pfeffer
Stuttgart Media University
`pfeffer@hdm-stuttgart.de`

- Motivation
- Instance-based matching
- Application to bibliographic data
- Evaluation
- Ongoing projects
- RDF Representation

Motivation

- Five regional library unions
 - Subject headings
 - Predominantly RSWK („Regeln für den Schlagwortkatalog“ - „Rules for the subject catalogue“) using a shared authority file
 - Classification systems
 - RVK (Regensburg Union Classification)
 - BK (Basic Classification)
 - DDC (Dewey Decimal Classification)
 - Various local classification systems
- Low proportion of indexed titles (25-30%)

- National library
 - Subject headings
 - Predominantly RSWK („Regeln für den Schlagwortkatalog“ - „Rules for the subject catalogue“) using a shared authority file
 - Classification systems
 - DDC (Dewey Decimal Classification)
 - Coarse categories
- DDC only for titles published since 2007
- Only „Reihe A“ (print trade publications) is fully indexed with RSWK

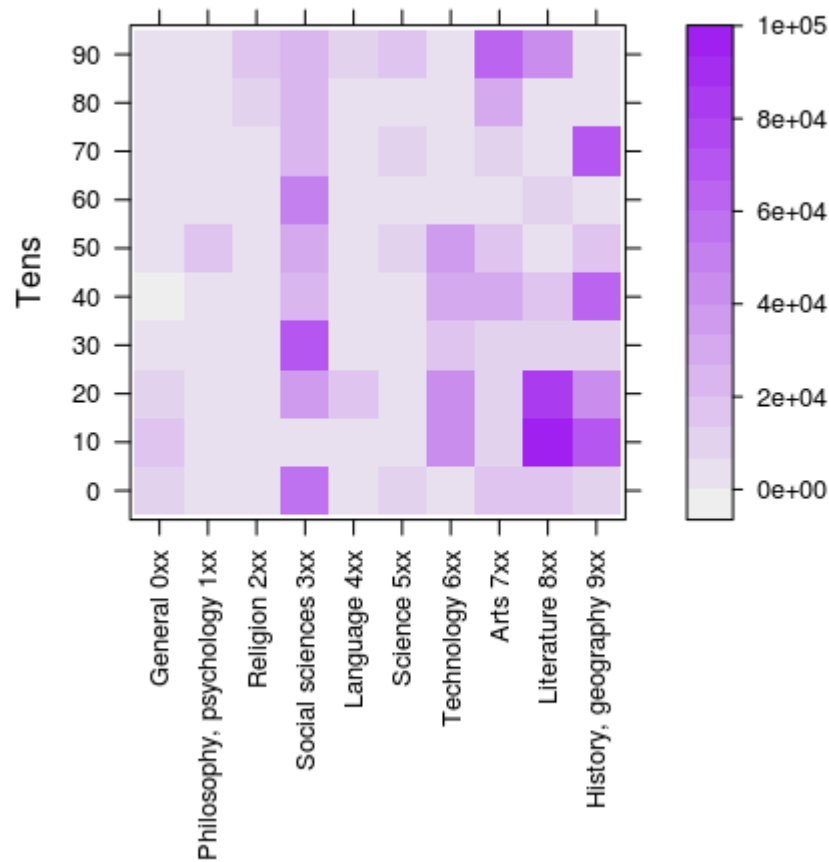
- Subject headings
 - Predominantly RSWK („Regeln für den Schlagwortkatalog“ - „Rules for the subject catalogue“) using a shared authority file
- Classification systems
 - BK since 2007
 - RVK in the Austrian library union catalogue

- Re-use existing indexing information
 - National level
 - BK is used mainly in northern Germany / Austria
 - RVK mainly in southern Germany
 - DDC mainly by the National Library
 - International level
 - Make RVK data more accessible to DDC users
 - Use DDC indexing information available from e.g. the Library of Congress

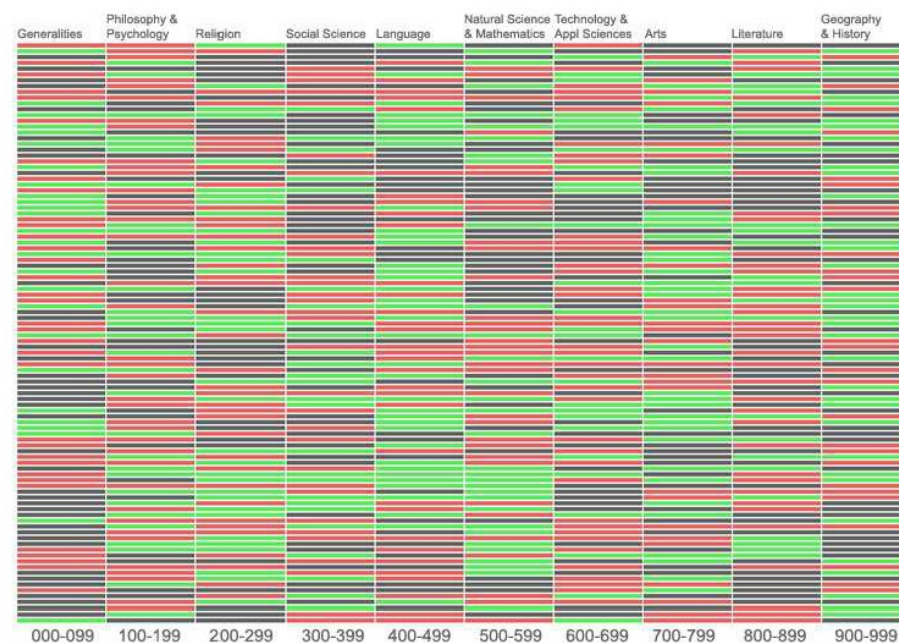
- Use of appropriate classification systems
 - Facetted search in resource discovery systems
 - Should be monohierarchical
 - Should have limited number of classes
 - DDC (first digits) or BK
 - Browsing of similar titles
 - Should be fine-grained
 - DDC (full) or RVK
- (Multi-lingual retrieval)

- Enable the use of existing tools and visualisations

TPL one-by-one dendrogram



DEWEY DECIMAL SYSTEM



Denton (2012) Hundreds

Legrady (2005)

Instance-based Matching

- Well-studied problem in computer science
- Several approaches
 - Based on the **descriptors**
 - Based on the **structure**
 - Based on the **manifestations (instances)**

- Entries in catalogues with multiple classifications

PPN:	369253876 Zitier 
Titel:	Dankbarkeit / Thomas Nisters
Verfasser:	Nisters, Thomas [1955-]   
Erschienen:	Würzburg : Königshausen u. Neumann, 2012
Umfang:	229 S. : graph. Darst. ; 235 mm x 155 mm
ISBN:	978-3-8260-4883-8 (Pb. : ca. EUR 39.00 (DE), ca. EUR 40.10 (AT)); 3-8260-4883-0
RVK-Notation:	CC 7250 INFO → Ähnliche Literatur
Sachgebiete:	Sachgruppe(n) DB (ab 2004) 100 DNB-DDC 179.9
Schlagwortfolge:	* Dankbarkeit  ; Philosophie  → Zum Register

- Assumptions
 - Classes with semantic overlap co-occur in instances
 - The more often these classes co-occur, the stronger the overlap
- Preparation
 - Extraction of all *pairs* of classifications from the data
 - Count of the extracted pairs


PPN: 31872197X [Zitier](#)
Titel: [Ermutigung zum unzeitgemäßen Leben](#) : ein kleines Brevier der Tugende und Werte / André Comte-Sponville. Dt. von Josef Winiger ...
Verfasser: [Comte-Sponville, André](#) [1952-]   
Beteiligt: [Winiger, Josef](#) ; [Volland, Nicola](#) ; [Pfau, Una](#)
Erschienen: Reinbek bei Hamburg : Rowohlt, 2010
Umfang: 414 S. ; 190 mm x 125 mm
Einheitssachtitel: [Petit traité des grandes vertus <dt.>](#)
Schriftenreihe: [Rororo ; 62599 : rororo-Sachbuch](#)
ISBN: 978-3-499-62599-2




RVK-Notation: [CC 7250](#) [INFO](#) | [CC 7200](#) [INFO](#) → [Ähnliche Literatur](#)

Sachgebiete: DDC [179.9](#)

Sachgruppe(n) DB (ab 2004) [100](#)

DNB-DDC [179.9](#)



Schlagwortfolge: *[Tugend](#)  → [Zum Register](#)

PPN: 369253876 [Zitier](#)
Titel: [Dankbarkeit](#) / Thomas Nisters
Verfasser: [Nisters, Thomas](#) [1955-]   
Erschienen: Würzburg : Königshausen u. Neumann, 2012
Umfang: 229 S. : graph. Darst. ; 235 mm x 155 mm
ISBN: 978-3-8260-4883-8 (Pb. : ca. EUR 39.00 (DE), ca. EUR 40.10 (AT)); 3-8260-4883-0

RVK-Notation: [CC 7250](#) [INFO](#) → [Ähnliche Literatur](#)

Sachgebiete: Sachgruppe(n) DB (ab 2004) [100](#)

DNB-DDC [179.9](#)

Schlagwortfolge: *[Dankbarkeit](#)  ; [Philosophie](#)  → [Zum Register](#)

■ Entry 1

- DDC: 179.9
- RVK: CC 7200
- RVK: CC 7250

■ Entry 2

- DDC: 179.9
- RVK: CC 7200

■ Pairs

- 179.9 / CC 7200
- 179.9 / CC 7250
- 179.9 / CC 7200

- Comparing solely absolute numbers is bad
 - Some classes are more often used than others
 - Number of pairs correlates with the number of entries that are classified using a given class
- Instead:
Use proportion of co-occurrence \leftrightarrow occurrence

$$\frac{|E_{c1} \cap E_{c2}|}{|E_{c1} \cup E_{c2}|}$$

number of entries with both classifications
divided by
number of entries with either classification
(Jaccard measure for overlap of sets)

- a and b are two classes from two classification systems A and B
 - The classes a and b only occur together
 - exact match
 - a only co-occurs with b , but b co-occurs with other classes from A
 - a is narrower concept than b
 - a co-occurs with several classes from B (including b)
 - a is wider concept than b
 - a and b do not co-occur
 - cannot infer that a and b are unrelated

- Pfeffer (2009)
 - Analysis of classification system structure and actual use
 - Locating classes that describe the same concept
 - Finding ways to improve existing mappings to RVK
 - Focus on RVK, using data from library union catalogues
 - Co-occurrence analysis
 - Results
 - High co-occurrence and close in the hierarchy:
→ classes are hard to assign properly
 - High co-occurrence and far in the hierarchy:
→ classes describe identical concepts
 - Mappings from RSWK to RVK could be augmented



- Isaac et.al. (2007)
 - Applied instance based matching to bibliographic data
 - Data from the National Library of the Netherlands
 - Mapping from a thesaurus to a classification system
 - Results
 - Generated mappings are quite good
 - More sophisticated measures than Jaccard do not lead to better mappings

Application to bibliographic data

■ Multiple editions

1. [Ermutigung zum unzeitgemäßen Leben : ein kleines Brevier der Tugenden und Werte](#)
Comte-Sponville, André. - Reinbek bei Hamburg : Rowohlt, 2010
(Rororo;62599 : rororo-Sachbuch)
-> Inhaltsverzeichnis
2. [Ermutigung zum unzeitgemäßen Leben : ein kleines Brevier der Tugenden und Werte](#)
Comte-Sponville, André. - 3. Aufl.. - Reinbek bei Hamburg : Rowohlt, 2004
(Rororo;60524 : rororo-Sachbuch)
3. [Ermutigung zum unzeitgemäßen Leben : ein kleines Brevier der Tugenden und Werte](#)
Comte-Sponville, André. - Reinbek bei Hamburg : Rowohlt, 1998
(Rororo;60524 : rororo-Sachbuch)
4. [Ermutigung zum unzeitgemäßen Leben : ein kleines Brevier der Tugenden und Werte](#)
Comte-Sponville, André. - 1. Aufl.. - Reinbek bei Hamburg : Rowohlt, 1996
-> Rezension

■ Multiple document types

- 
1. [Web X.0 : Erfolgreiches Webdesign und professionelle Webkonzepte; Gestaltungsstrategien, stationäre und mobile Medien \[Elektronische Ressource\]](#)
/ Stapelkamp, Torsten. - Berlin, Heidelberg : Springer Berlin Heidelberg, 2010
Link zum Volltext: [Elektronische Ressource: Zugang beim Produzenten](#)
 2.  [Web X.0 : erfolgreiches Webdesign und professionelle Webkonzepte; Gestaltungsstrategien, stationäre und mobile Medien](#)
Stapelkamp, Torsten. - Berlin : Springer, 2010
-> Inhaltsverzeichnis

- Skewed data
 - Multiple editions → More pairs
 - Some co-occurrences could appear stronger than others
- Solution: Pre-clustering individual titles on the „work“ level
 - Increases chance for instances with more than one classifications
 - Each cluster contributes only once
 - Allows using absolute co-occurrence numbers
 - Cut-off for small numbers
 - Ranking of competing matches

- Pfeffer (2013)
 - Matching bibliographic records
 - Based on author, title and uniform title
 - (as well as information on title changes)
 - Matches any edition and revision of a work
 - Including translations
 - Merge match sets → Discrete clusters
 - Consolidating indexing information
 - For indexing purposes, the differences between editions and revisions are irrelevant
 - Subject headings and classifications are shared between all members of a cluster

Evaluation

- Existing (partial) mappings can be used as a basis for evaluation
 - „Gold standard“
- Comparison of automatic and manual mapping
 - Recall: Are all the mappings found?
 - Precision: Are all found mappings correct?
- Analysis of additional links
 - Maybe the gold standard can be improved?

Ongoing projects

- Bibliographic data
 - German library union catalogues
 - German National Library catalogue
 - Austrian National Library catalogue
 - British national bibliography

- Gold standards
 - Partial mappings BK ↔ RVK

- RVK → BK
 - Gold standard exists
 - BK well suited for faceted retrieval
 - RVK has largest proportion of classified titles

- RVK ↔ DDC
 - Enable data sharing between the German National Library and the RVK-using libraries

- Not limited to classification systems
 - See Pfeffer (2009) and Wang et.al. (2009)

- Import and mapping of MAB2 and MARC data
- Clustering
 - Generation of keys for the match process
 - Matching and clustering
 - Consolidation of indexing and classification information
- Statistics
 - Co-occurrence counts
 - Jaccard measure
- Output
 - Full mappings

- All steps implemented as a prototype
 - Perl scripts
 - File-based data and indexes
- Current development
 - Still Perl scripts (but better documented)
 - All data is accumulated in a document store
 - MongoDB
- Further plan: Porting to MetaFacture framework

RDF representation

- DDC has been published as Linked Data
- RVK has **not** been published as Linked Data
 - There is no versioning and no stable identifiers
 - A project to fix this and to publish RVK as Linked Data has been cancelled by the university library of Regensburg
- BK has not been published as Linked Data
 - There is authority data in the GVK union catalogue

→ One would have to create temporary URIs for the RVK and BK classes

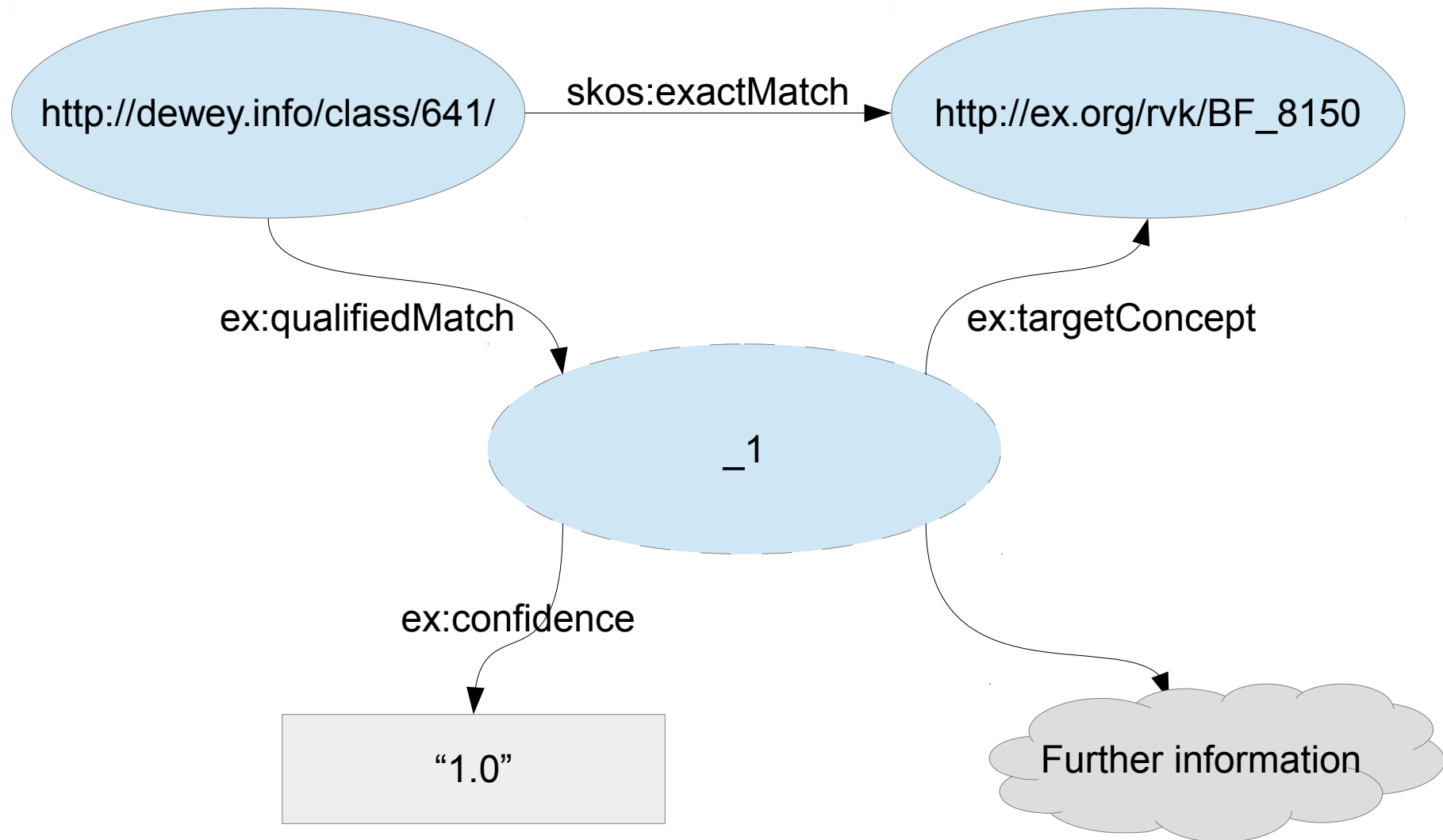
- SKOS offers
 - `skos:mappingRelation`
 - `skos:closeMatch`
 - `skos:exactMatch`
 - `skos:broadMatch`
 - `skos:narrowMatch`
 - `skos:relatedMatch`

- 1:n-Relationships
 - List of classes that are all narrow matches
 - Or: *A combination* of classes is a (near) exact match
- Qualified mappings
 - Express the confidence of the proposed match
 - Allow applications to optimize for precision or recall

- RDF relations cannot be qualified



- So an intermediate node is used



- Mappings between classification systems are an important means for interoperability and sharing of classification information and tools
- Simple statistics on the existing data in catalogues can generate candidates for matches between individual classes
- Where manually created mappings exist, they can be used to evaluate the algorithmic results
- Mappings can be expressed in SKOS terms
 - To qualify the mappings further, intermediate nodes need to be introduced
 - There is no standard for this yet

Thank you for listening.

Slides available online
<http://www.slideshare.net/MagnusPfeffer/>

This work is licensed under a Creative Commons
[Attribution-ShareAlike 3.0 Unported License](https://creativecommons.org/licenses/by-sa/3.0/).



- Denton, W. (2012). [On dentographs](#), a new method of visualizing library collections. In: *Code4Lib*.
- Isaac, A., Van Der Meij, L., Schlobach, S. and Wang, S. (2007). [An empirical study of instance-based ontology matching](#). In: *The Semantic Web* (pp. 253-266). Springer Berlin Heidelberg.
- Legrady, G. (2005). [Making visible the invisible](#). Seattle Library Data Flow Visualization. In: *Digital Culture and Heritage. Proceedings of ICHIM05 Sept, 21-23*.
- Pfeffer, M. (2009). Äquivalenzklassen – Alle Doppelstellen der RVK finden. Presentation given at the Librarian Workshop of the 33rd Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications (GfKI).
- Pfeffer, M. (2013). Using clustering across union catalogues to enrich entries with indexing information. In: *Data Analysis, Machine Learning and Knowledge Discovery - Proceedings of the 36th Annual Conference of the Gesellschaft für Klassifikation e. V.* Springer Berlin Heidelberg.
- Wang, S., Isaac, A., Schopman, B., Schlobach, S. and Van Der Meij, L. (2009). [Matching multi-lingual subject vocabularies](#). In: *Research and Advanced Technology for Digital Libraries* (pp. 125-137). Springer Berlin Heidelberg.