Sebastian Gabler*

Thesauri – a Toolbox for Information Retrieval

https://doi.org/10.1515/bfp-2023-0003

Abstract: Thesauri are established tools for subject indexing. Through recent technological evolution and the rise of artificial intelligence they have become more relevant by their ability to deliver explainable results for computer-aided indexing- and concordance work with other data sets and models, and for data validation. Based on existing own research for a master thesis we expand on the aspect of quality assurance in library catalogs, following selected examples.

Keywords: Subject indexing, thesaurus, information retrieval

Thesauri als Werkzeuge für Information Retrieval

Zusammenfassung: Thesauri sind etablierte Instrumente der bibliothekarischen Sacherschließung. Durch die jüngste technologische Entwicklung und das Aufkommen künstlicher Intelligenz haben sie an Bedeutung gewonnen, da sie in der Lage sind, erklärbare Ergebnisse für die computergestützte Erschließungs- und Konkordanzarbeit mit anderen Datensätzen und Modellen sowie für die Datenvalidierung zu liefern. Ausgehend von bestehenden eigenen Recherchen für eine Masterarbeit wird der Aspekt der Qualitätssicherung in Bibliothekskatalogen anhand ausgewählter Beispiele vertieft.

Schlüsselwörter: Sacherschließung, Thesaurus, Information Retrieval

1 Introduction

The creation of a thesaurus starts with the definition of its domain. The domain of the 1856 publication of Peter Mark Roget "Thesaurus of English Words and Phrases so as to facilitate the expression of ideas"[1] is the English language.

The reasons why Roget's Thesaurus still is iconic for information professionals result from his sparse use of classes and hierarchy. Organizing approximately 450 000 terms into only six primary classes, omitting the extensive use of hierarchy, Roget displays the full extent of linguistic challenges, including synonyms, antonyms, meronyms (whole/partial relations) and that of equivocation (homonyms and polysemes).[2] Despite its long history, it continues to inspire people and still is used as a blueprint for current knowledge representation standards and indexing methods. We even have adopted the word *thesaurus* as a coined term for a type of controlled vocabulary, totally changing it from its original Latin/Greek meaning of "treasury" or "warehouse".[3]

This paper builds on the master thesis "Vergabe von DDC-Sachgruppen mittels eines Schlagwort-Thesaurus" presenting a method for the assignment of subject categories for scientific publications by automated extraction of metadata from natural language.[4]

To that end, a text is indexed against the subject headings of Gemeinsame Normdatei (GND), using a computational linguistic processor. The documents are then classified, applying transitive closures for the indexed keywords along the hierarchy provided by Dewey Decimal Classification (DDC), grouping them per DDC Sachgruppen (subject categories, SC)[5] with an aggregated and weighted relevancy score.[6]

The method, also presented at the 6. Workshop Computerunterstützte Inhaltserschließung,[7] is suitable for a wide range of indexing workflows.

The ranking relevancy score of the retrieved subject categories may be used for automated or manual assignment. Thus, it is specifically suitable for incremental and collaborative indexing approaches. The approach may also serve as a complementary method for validation of existing automated indexing results, specifically those from machine learning. This aspect will be examined in detail in section 3 of this paper.

1 Roget (1856).

***Kontaktperson: Sebastian Gabler, MSc.**,
sebastian.gabler@sebastian-gabler.eu.
https://orcid.org/0000-0002-8607-0952

2 Roget's Thesaurus has been a steady companion of the author for writing texts in English as a foreign language, starting with writing a preparatory paper in the senior year of high school.

3 https://en.wikipedia.org/w/index.php?title=Thesaurus&oldid=1132315452#Etymology.

4 Gabler (2021).

5 https://www.dnb.de/DE/Professionell/DDC-Deutsch/DDCinDNB/ddcindnb_node.html#doc259608bodyText1.

6 Gabler (2021) 72 ff.

7 https://wiki.dnb.de/pages/viewpage.action?pageId=252121510&preview=/252121510/266437041/DA3%20Workshop-Gabler.pdf.

Automated indexing has been established for subject cataloguing for more than a decade. Today, in most cases, AI-based methods, especially those employing machine learning (ML), are being used.[8] These methods are notorious for the so-called "cold start problem": a supervised or unsupervised learning process first establishes the challenge to be solved by the machine. Often, this is done using annotated reference corpora per knowledge domain.

While universal classification systems such as GND and DDC cover all fields of literature, universal reference corpora often may not be available as we depend on the existence sizeable sets of pre-classified publications for each subject category.

Another disadvantage of ML-based methods is their basic opaqueness. Performance may be judged based on predictions made by a black box against the verdict of subject matter experts. How the machine arrived at its decisions and whether it possibly made an error, is hardly ever comprehensible.

An alternative method is the use of a description logic, especially in the form of a thesaurus, in combination with statistically weighted indexing of the full text. Here, too, extensive preliminary work must be done, specifically the creation of the thesaurus which may present a challenge, specifically if the thesaurus represents all domains of world knowledge. This may contribute to the fact that approach is less frequently described in literature.[9] However, the indexing procedure based on description logic is quite effective and has characteristics complementary to AI-based methods. Decisions supported or made by a symbolic description logic are always comprehensible, including an indication of failure and success immediately derived from the underlying ruleset.

When introducing and operating automated indexing aids one of the main challenges is validation of results. A systematic methodology has been established by Korjalka Golub at the National Library of Sweden.[10] Golub formalizes a multi-stage procedure for this purpose as follows:

- Evaluation of indexing by comparison with a gold standard.
- Evaluation of the quality of computer-assisted indexing in the context of an indexation workflow.
- Assessment of indexing quality through analysis of retrieval performance.

Using this validation approach on the proposed method, we could show that the thesaurus-based approach was able to produce good to excellent results for research papers in selected subject categories, retrieved from the online catalog of the German National Library (DNB). The performance was comparable with the results of previous machine-aided indexing methods, while using a complementary approach.[11]

2 Definitions and theoretical background

Controlled vocabularies have been traditional companions of subject catalogers for a long time. With the rise of information technology, specifically with the introduction of the Semantic Web, taxonomies, classification systems, thesauri, and other knowledge organization systems (KOS) have become valuable AI tools for information professionals. Encoding librarian cataloging standards into machine-readable models based on web standards enable smart automation, following the underlying mathematical models, thus taking the existing systems into a new era.

2.1 Standards for knowledge organization systems

Following ISO 25964-1 (2011), Information and documentation – Thesauri and interoperability with other vocabularies,[12] a thesaurus is a "set of selected concepts represented primarily by preferred terms, formally organized so that paradigmatic relationships between the concepts are made explicit, and the preferred terms are accompanied by lead-in entries for synonyms or quasi-synonyms". The comprehensive set of concepts describes a knowledge domain formally.

There are small thesauri describing niche domains, such as the Graz thesaurus for object classification in the domain of cultural heritage (229 concepts)[13] and thesauri aiming at the general published knowledge of German language, such as Gemeinsame Normdatei (GND) of the national libraries of Austria, Germany and Switzerland (many millions of concepts, when including persons, corporate

8 Golub (2019).
9 https://www.isko.org/cyclo/automatic#3.3, accessed 2023-11-01.
10 Golub et al. (2016).
11 Gabler (2021).
12 ISO 25964-1 (2011).
13 http://gams.uni-graz.at/archive/objects/o:oth/methods/sdef:SKOS/get.

bodies and geographic names).[14] In-between the extremes illustrated here, a range of thesauri of all sizes exists.

Thesauri overlap with other forms of KOS, i. e., authority lists, terminologies, classification schemes and ontologies. ISO 25964 is built on a conceptual approach applicable to all kinds of KOS. ISO 25964:2013-2 (part two of the aforementioned standard) mainly delimits them by their primary representation of concepts. I.e., preferred terms representing thesaurus concepts, category labels or notations used in taxonomies, and notations for classification schemes.[15]

Other authors attempt to classify KOS mainly by the expressivity of knowledge representation. Mainly, expressivity is given by the completeness of context, as Pellegrini and Blumauer put it in ascending order for flat authority lists, monohierarchic taxonomies, polyhierarchical thesauri and comprehensive ontologies.[16]

Practically, the complexity of a KOS will typically grow along the life cycle. Thus, a KOS classification following ISO is more stable and very useful when we are to create crosswalks between different indexing system types.

Thesauri and other KOS today are linked globally, as seen with the Linked Open Data Cloud (LOD).[17] Specific LOD crosswalks that also exist for GND make it possible to evaluate the international context of a subject or discipline, using information that is explicitly encoded and immediately available, using web standards.

We may also employ several conceptual representations contained in an indexing system, not only the primary representation. The convergence of different types of indexing systems may make them specifically attractive. I.e., while the primary representation of DDC are numbers (notations) it also includes a comprehensive vocabulary that makes it also useful as a facet system for search or even for keyword search.

2.2 Mathematical models

Eventually, all automated indexing methods are following mathematical models. In the field of Artificial Intelligence (AI), thesauri and taxonomies are part of the sub-domain of Symbolic AI. “Symbolic” in this context refers to concepts and context represented in the form of human-readable symbols, that is information, whereas sub symbolic AI is directly processing the underlying data for patterns and using complex algorithms.[18] Here, thesauri and taxonomies provide knowledge representation and allow reasoning and inference based on formal logic, specifically provided by context.

Taxonomies are following the paradigm of subsumptive containment hierarchy

$$x \subsetneq y,$$

whereas we can say classes (*x, y)* are different from each other, class *y* contains class *x*, and *x* is more specific than *y*.[19]

Thesauri are directed graphs, containing both hierarchical and associative relations.[20] A more specific form is the directed, acyclic graph (DAG), which is guaranteed to be loop-free. This mathematical model allows polyhierarchy (i. e. (*b*, *c)* are generic concepts of *d*) but it has no cyclic relations, such as *d* being related to *c*. The absence of loops in a thesaurus allows for a more straight-forward transitive closure and aggregation.[21]

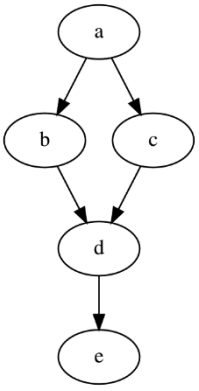

Fig. 1: Directed, acyclic graph[22]

Transitive closure (let *a* be a generic concept of *b*, *b* be a generic concept of *d*, then *a* is also a transitive generic concept of *d*) is possible in both KOS types, taxonomy, and thesaurus.

A thesaurus also offers closures along non-hierarchical, associative relations. For classification, associative results are mainly useful for disambiguation.

14 https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html.

15 ISO 25964-2 (2013).

16 Pellegrini and Blumauer (2006).

17 https://lod-cloud.net/.

18 C.f. https://towardsdatascience.com/symbolic-vs-subsymbolic-ai-paradigms-for-ai-explainability-6e3982c6948a, accessed 2023-01-14.

19 https://en.wikipedia.org/w/index.php?title=Hierarchy&oldid=1128328391#Subsumptive_containment_hierarchy.

20 C.f. de Jesus et al. (2004).

21 Gabler (2021) 53 ff.

22 https://commons.wikimedia.org/wiki/File:Tred-Gprime.svg License: Public Domain.

2.3 Web technology standards

The World Wide Web Consortium (W3C) has made significant contributions to this with the Simple Knowledge Organization System (SKOS)[23] Both, SKOS and ISO follow a concept-based approach leading to a convergence of verbal and classifying indexing methods. SKOS is an application of the Resource Description Framework (RDF), a universal, technical method for describing arbitrary things on the Web.[24]

When expressed in RDF, any concept is represented by its International Resource Identifier (IRI). IRI-carrying resources can be identified, accessed, and processed via standardized transmission and directory protocols (Web technology), linking natural language and machine language.[25]

SKOS concepts (terms) carry the naming of things (preferred naming and synonyms), thus enabling indexing and faceting of unstructured texts with the thesaurus. A thing in SKOS represents an abstract unit of thought. Using SKOS concepts, we can describe real and virtual objects or individuals and abstract classes. While the attribute "simple" may indicate otherwise, SKOS is an instance of an ontology following the Web Ontology Language (OWL) standard, a comprehensive knowledge representation based on description logic. "Simple" means the reduction of hierarchical and associative relations to their fundamental patterns which are also the relations most frequently used in real-world thesauri.

In the field of library indexing, extensive use has been made of SKOS: With the Library of Congress Subject Headings, Standard Thesaurus Wirtschaft (STW), the French RAMEAU and the Dewey Decimal Classification, essential verbal and classification indexing systems are available as SKOS datasets. GND subject headings are as well organized as an ontology based on SKOS classes and attributes.

We establish that GND subject headings are a valuable mathematical representation of RWSK: given the concepts (subject headings) are correctly applied by natural language processing, we can index natural language automatically and reason along the context using the indexate for various known applications, such as categorization, verbal indexing, and faceting, but also for novel applications as collaborative indexing, data validation and repair.

The underlying master thesis is based on description logic that exists today in GND and DDC. Employing the defined overlay of these systems encoded as DDC notations yields metadata eventually corresponding to Dewey classes. The overlay may be exploited by materialization[26] or also by inference, with equivalent results.

3 Validating existing classifications

Following standing indexing policies, electronic resources typically have subject categories assigned by an automated process only. This has been particularly the case for Series O of the *Deutsche Nationalbibliografie* for over a decade[27] but also for related resources such as publication databases like PubMed.[28] Currently, this is typically done using binary classifiers employing the machine learning approach, such as Support Vector Machine.[29] While these methods are successful assigning classifications consistently, subject categories assigned by artificial intelligence, and even those manually assigned, are not always correct.

During the collection of a pre-labelled reference corpus for my master thesis from the online catalog of the German National Library,[30] it was inevitable to stumble across several cases with wrong subject categories assigned. In the following section, we will discuss two of the items found and expand on the possibilities of data validation and data repair along these examples.

Note: Considering the rather cursory nature of the present endeavor, the indexing results of sample resources cannot claim to be comprehensive or fully authoritative. Still, the insights gained from its samplings demonstrate characteristics of the method based on existing research and are supported by the validity of the underlying models.

3.1 Validation scenarios and possible resolution methods

Validation includes several stages, not limited to:

- Establishing key performance indicators, specifically expected values for index terms or classes,
- Identification of candidates i. e., by pre-selection or by batch indexing,

23 https://www.w3.org/TR/skos-primer/.

24 https://de.wikipedia.org/w/index.php?title=Resource_Description_Framework&oldid=209155194.

25 C.f. https://www.rfc-editor.org/rfc/rfc3987, accessed on 2022-12-30.

26 Gabler (2021) chapter 5.

27 C.f. DNB (2019).

28 NLM Technical Bulletin 443: https://www.nlm.nih.gov/pubs/techbull/nd21/nd21_medline_2022.html, accessed on 2023-01-09.

29 Mödden (2012), Nentidis et al. (2022).

30 Gabler (2021) 94 ff.

– Comparison with expected values,
– Evaluation of results,
– data repair.

For each of these stages, we can apply automation, depending on the defined targets. Often, we may automatically identify a candidate, that is a document having a potentially incorrect indexate, but we cannot fix it immediately.

3.2 Automated error detection and repair

Error sources may be as trivial as typos, such as the wrongly assigned subject category Mathematics (**5**10) of "Untersuchungen zum Einfluss des Leukozyten-Inhibitions-Moduls auf die posthämorrhagische Inflammation und Gewebehypoxie im Tiermodell"[31] that should rather have been classified as a dissertation in Medicine (**6**10). While the transitive closure of the (correct) DDC Short Number 617.2 in the indexate already provides a hint to the error, using the annotation thesaurus confirms minimal relevance of the catalogued subject category 510 automatically. (1/20 of the score, relatively to the first ranking category, rank 15).[32]

In the case instanced, we can validate that both, automatic detection of the error and correct classification are very reliable because we have previously validated that both predicted and found classes (510 and 610) are determined with high confidence within a gold standard. In the research for the master thesis, mathematical and medical publications were among the 14 target subject categories for validation, both with excellent results.

With the Mean Recursive Rank value at 0,96, only 4 % of the tested mathematical publications did not come back with the predicted class in the first position. No single document in Mathematics came back worse than third position. For Medicine, only 2 % were predicted below third rank.[33] Additionally, we can count the number of overlapping concepts between the two subject categories, rendering zero results among SC 510 and SC 610 for the thesaurus used.

Hence, we can conclude here that both predictions about the wrong subject category previously assigned, and the correct subject category predicted are correct. We can also say that validating documents in Mathematics and Medicine will probably render good to excellent results by this method, as the thesis has shown for many other relevant subject categories with high publication frequency.

3.3 Automated error detection, manual repair

With the same conviction however, we state it may be not that clear for all examined scenarios and subject categories.

In a specifically interesting case, we can determine with high confidence the existing subject category is probably wrong. However, challenges with the indexing rules make it difficult for the automatic indexer to predict the correct subject category for an entire class of publications, here for business informatics.

The examined publication "Branchenspezifische IT-Innovationssysteme: von der Analyse zur Intervention; am Beispiel des IT-Innovationssystems für Krankenhäuser in Deutschland" was and still is indexed incorrectly with SC 360 (Social problems and services).[34]

3.3.1 Error detection

With some confidence, we assert that this subject category assignment was wrong. We cannot judge the source of the error, specifically if the subject category was manually assigned or automatically without looking at the source data. However, SC 360 as the only class assigned should raise concerns with the aboutness of the text: the paper is investigating special IT systems for hospitals. By transitive closure, SC 360 would indeed be the correct class for a paper about hospitals (Dewey number 362.11).[35] A short autopsy of the publication however reveals the publication is about IT systems, written by a computer scientist and reviewed by a computer scientist, with hospitals being a side aspect – even if a relevant one.[36]

Following RWSK §§ 6 and 13.4, the categorization of the document under SC 360 consequently is wrong.[37] This is also supported by the ranking of the predicted classes with our indexing machine, returning SC 360 in position 15 (see Figure 2).

The thesaurus was helpful for pointing at a possible error automatically. Because it indexes the full text, the method is reliable and robust for the detection of erroneously indexed side aspects.

31 https://portal.dnb.de/opac/simpleSearch?query=idn%253D100748247, still incorrectly classified on 2023-01-09.
32 Gabler (2021) 95.
33 Gabler (2021) 84 ff.
34 https://portal.dnb.de/opac/simpleSearch?query=idn%253D1021499684, retrieved 2022-12-28.
35 https://d-nb.info/gnd/4032786-3.
36 For the challenges of indexation of side aspects see Gabler (2021) 15, seq. Gödert et al. (2014) 149 ff.
37 DNB (2017).

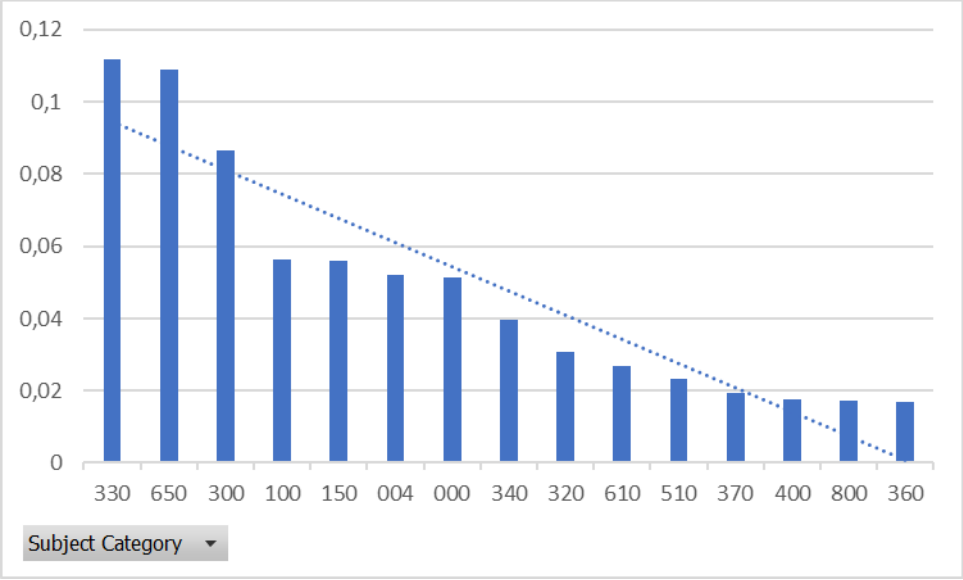

Fig. 2: Top 15 result for document IDN 253D100748247

3.3.2 Repair attempt

The question which subject category would be desirable for this paper however cannot be answered by the thesaurus, as-is. This has several reasons which seem to be relevant for the specific discipline of business informatics (Wirtschaftsinformatik) and informatics in general.

A possible subject category would be 004, Informatics. Informatics only comes up for this document as a category in 6th rank.

Figure 2 displays the detailed top-15 ranks returned from the indexer, in descending score rank. Apparently, there are no strong statistical indications for a specific single subject category.

While the transitive closure for Wirtschaftsinformatik in GND following the Dewey number 330.0285 indeed would be SC 330, there is no substantial distance to the next ranking SC 650 – Management from the automatic indexing run.

At this point, we can already state the quantitative result is not sufficiently significant to allow taking a decision only from the predictions of the indexer. Distribution of classes returned is a strong indicator of success or failure which many machine learning systems, specifically binary classifiers, cannot provide.

The underlying issue however does not seem to be an accidental mismatch of text and vocabulary. The results are rather exemplary and show a specific anomaly of the classification of the discipline of informatics, specifically business informatics in GND.

First, we observe a surprisingly small result set for business informatics i. e., by doing a DDC-search for number 330.0285 from Series O in the DNB online catalog.[38] This search retrieves 65 results, only. From the publications in this result set, those having subject category assigned are by majority indexed with SC 330. Based on a suspiciously small result set, this is not a strong indicator; hence, we deem those classes as non-representative. It is therefore doubtful GND subject headings really establish business informatics publications as a topic under SC 330, Economy, in a sustainable manner.

Alas, business informatics also seems to have its challenges under SC 004, Informatics. This is because of incompleteness and bias in the specific set of concepts that close into this category, as we will later see. This has already shown in the master thesis, where informatics was among the relatively low-performing categories (P = 0.72, R = 0.62).[39]

Looking at precedent work, the ML-based indexing system used in PETRUS seems to have similar challenges with Informatics. With recall below 0,6 and precision even below 0,4, the indexing performance for that subject category was at the bottom end of that study.[40] The only indicator we have for the practical success of PETRUS-driven classification of business informatics publications are the findings in the DNB online catalog, which are not many, as mentioned.

38 https://portal.dnb.de/opac/showShortList?currentPosition=4&currentResultId=dcs%3D330.0285%26any%26online.
39 Gabler (2021).
40 Uhlmann (2013).

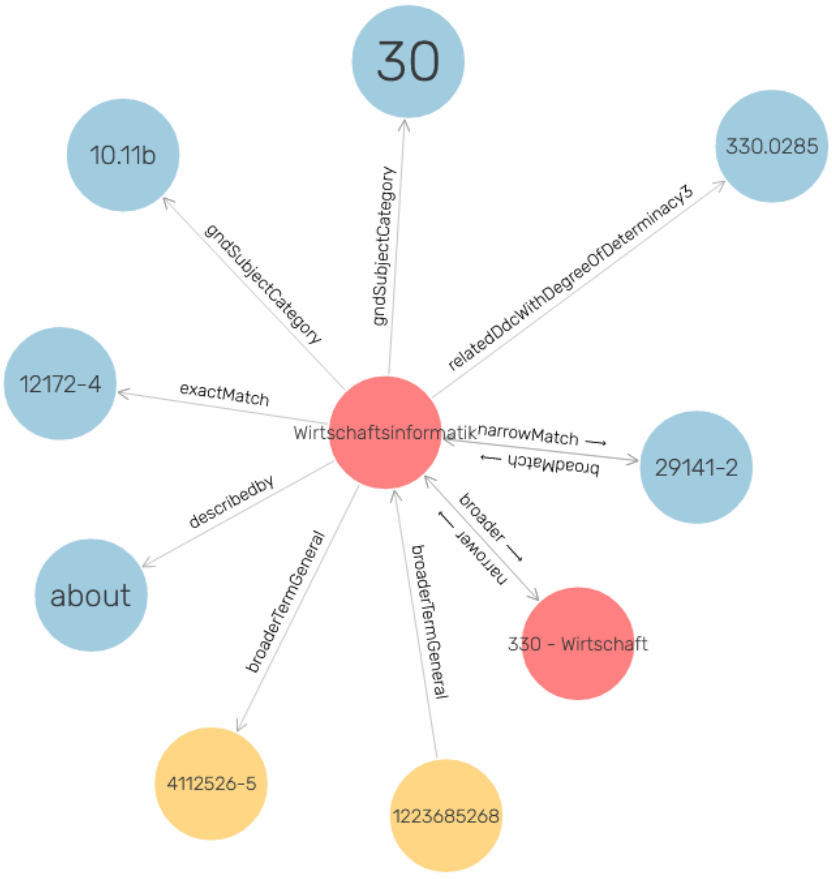

Fig. 3: Wirtschaftsinformatik with direct neighbors in GND

3.3.2.1 Conceptual context in the thesaurus

With reference to the GND authority data, the heading Wirtschaftsinformatik is a leaf concept. That means, it has no descendants or siblings by the GND hierarchy and context. Consequently, we have no additional vocabulary to integrate, producing structural sparseness in classification.

Within GND, Wirtschaftsinformatik has three broader concepts. It has associative links to other authority data in STW which themselves point towards informatics, not to economy.[41]

Both, STW and GND concepts are mapped to Wikidata, a KOS frequently used for semantic AI applications. While the concept has the English label “business informatics” in the related Wikidata entry,[42] this label is not only missing in the GND data itself; most prominently we can state the concept has no direct equivalent in relevant national or international authority data.[43] For an internationally relevant discipline with active publication records this is not expected as it will cause issues during information retrieval.

Moreover, we can observe a conflict between the GND systematic and DDC notation as the old systematic is pointing at classes 10.11b (Mathematische Methoden) and 30 (Informatik). We also observe a parent concept in GND (applied informatics) has no notations in Economy but rather in Management and Informatics.

The Dewey number for Wirtschaftsinformatik in GND is a concatenation of class 330 and a generic notation from auxiliary table T1-0285, application of computers in a specific field.[44] The literal translation of the corresponding verbal description of an activity would be “computer application in economy”. From an aboutness perspective, the question remains if the primary aspect of Wirtschaftsinformatik means business benefits of IT, or if it is the application of information technology in the economy. This is a question that has no trivial answer.

41 http://zbw.eu/stw/version/latest/descriptor/12172-4/about.

42 http://www.wikidata.org/entity/Q126552.

43 The best match is a skos:closeMatch to an equivocal concept in STW, see http://zbw.eu/stw/descriptor/12172-4.

44 https://www.oclc.org/content/dam/oclc/webdewey/help/table-1.pdf. For some reason, numbers concatenated with the same auxiliary extension cannot be browsed in WebDewey or BASE at the time of writing. Apart from the observations made in the master thesis with the auxiliary tables being orthogonally ordered to the main table hierarchies, we cannot explain this at this time. In any case, this has detrimental effects for findability.

Among the guiding principles assigning Dewey numbers to GND subject headings are focused and comprehensive mapping.[45] While business informatics contains much, if not all, informatics in the specific context of business, business is not (only) economy but also management. So, the question would be why SC 330 was the preferred parent for business informatics, and if that approach is focused and comprehensive for the topic.

Taking a systematic look from the broader term applied informatics,[46] we find that it is divided into 15 disciplines, one of them being business informatics. Among these siblings, the majority has multiple notations, often with a transitive closure into SC 004. Only two of them expose the same pattern as business informatics number, that is the composition using the auxiliary Dewey table. (Rechtsinformatik and Umweltinformatik).[47]

While we observe that WebDewey does not provide a browse entry under 330.0285, WebGND allows browsing that position of the Dewey hierarchy.[48] Hence, we can count the descendants and siblings, arriving at a total count of seven subject headings closing into the discipline, all of them bearing notations into SC 330, if any. So, again, we are confronted with a focused yet not comprehensive vocabulary.

Hence, also from a bottom-up approach we cannot explain the single assigned notation of business informatics into economy, analyzing related subject headings and the related vocabulary.

While we observe that the notation approach was focused, it certainly was not comprehensive. Thus, we state a sole DDC notation for the concept below 330 is maybe questionable and eve more that it was provided with the second-highest degree of determinacy (3 of 4), which should be "topics that are nearly coextensive with the full meaning of a class or cover more than half of the content of the class approximate the whole of the DDC class".[49]

3.3.2.2 Context of the most relevant subject heading

Often, a single subject heading describes the topic of a publication. Therefore, it is interesting to look at the top-ranked items here as well. The most relevant subject heading in the document was "Innovation"[50] with 412 occurrences in the document.[51]

The heading has transitive closures into SC 330 (one relation) and SC 650 (three relations). All DDC notations have been provided with a rather low determinacy, 2, which was the lowest value used for the indexer.[52] For a concept with rather generic meaning, this appears to be adequate.

While the indexer distributes the relevance equally between the two parent categories, the prototype does not account for the determinacy score and not for the weight provided by the number of relations (3:1 for SC 650).

This clearly is a weakness of the machine that should be improved. An extrapolated calculation reveals that SC 650 would have come up in top rank for the document, if only the number of relations would be considered.

However, also from this approach, the document would still not land anywhere below SC 330, but rather in SC 650.

3.3.2.3 Relation of Informatics to other subject categories

Taking a specific top-down approach is not possible for business informatics because of the lacking search entry under SC 330 (or SC 650) in DDC. Hence, we do not find any facet values that would point us to business informatics. However, the related subject category 004 has a sizeable vocabulary of 744 concepts. Overlaps can be established, using SPARQL queries into the transient closures to the parent subject categories.

Table 1 displays the quantity and distribution of the conceptual overlap of subject category 004, Informatics. We display the top 6 target categories with overlap to the source category, the two bottom ranks (whereas the penultimate row is the predicted subject category by the indexer for the examined document), the aggregated number of concepts with at least two parent categories, and the resulting degree of overlap and percentual distribution of the total overlap to all other categories.

Obviously, the vocabulary of informatics has significant overlap with other categories. This has been found for other subject categories and per se does not impair correct classification.[53]

Remarkable here however is the obvious bias of the overlap: with the top two overlapping categories being electrical engineering and industrial manufacturing, the vocabulary for Informatics is apparently quite a bit "hardware-biased". Only two concepts overlapping with business related topics clearly show business informatics are detached from the informatics, with almost no chance for any full text to fall into that category by hierarchical aggregation, eliminating chances of successful post-coordination.

45 Gödert et al. (2014) 108 ff.
46 https://d-nb.info/gnd/4112526-5.
47 https://d-nb.info/gnd/4762396-2; https://d-nb.info/gnd/4048802-0.
48 http://gnd.eurospider.com/s?table=ddc&node=33002855.
49 Gödert et al. (2014) 140.
50 https://d-nb.info/gnd/4027089-0.
51 The normalized relevance of the top item is always 100. Position and frequency are used for weighting.
52 Gabler (2021).
53 Gabler (2021) 67 f.

Tab. 1: Structure of conceptual overlap, Informatics in GND

Rank	Target Subject Category	# Overlapping concepts in target category	Percentage of 004 concepts (744)	Percentage of total overlap (1129)
1	621.3 – Elektrotechnik, Elektronik	573	77 %	51 %
2	670 – Industrielle und handwerkliche Fertigung	100	13 %	9 %
3	770 – Fotografie, Video, Computerkunst	84	11 %	7 %
4	020 – Bibliotheks- und Informations-wissenschaft	59	8 %	5 %
5	510 – Mathematik	56	8 %	5 %
6	650 – Management	49	7 %	4 %
...	...	...	...	...
23	330 – Wirtschaft	2	0 %	0 %
24	340 – Recht	2	0 %	0 %
	Total	1 129	152 %	100 %

For the top categories found for the document (SC 330 and SC 650), we can state that there is significantly more overlap of informatics with management, compared with that for economics.

Conversely, economics and informatics have quantitatively no overlap with economics following GND-DDC notations, raising the question if the focus of classifying business informatics into Economy was on target, or if Management would have been a better choice. Alas, there is no German term that allows to connect information technology and management by morphology as there is with business informatics in English.

3.3.2.4 Publication activity in predicted categories

Another approach to validate the approach is investigating related publication activity.

The publication activity per Dewey class can be estimated i. e., by doing a respective query in a database for scientific publications. With results as sparse in the DNB catalog as we previously found, we need an alternative base line. Bielefeld Academic Search Engine (BASE) is an obvious candidate for this task, as it is indexed using DDC, employing machine learning. BASE is indexing regional and international publications with an emphasis for Open Access.[54]

As BASE does not employ DNB subject categories but top-level DDC classes, we must combine the Dewey number range covering DNB-SC 004 for equivalent characteristics. This should be an aggregation of result sets for publications classified within the ranges under Dewey numbers of 004–006.[55] SC 330 and 650 are sufficiently consistent with the equivocal Dewey number ranges, hence we can use their results directly.

Finding publications on business informatics indexed with SC 330 is a bit tricky: searching for DDC number 330.285 renders no results in BASE. Browsing publications under DDC number 330 for business informatics papers initially is like finding the needle in the haystack.

A first approximation can be done by combining the query with a characteristic search string – “computer” was an obvious choice. As we can do the same for SC 650, we can retrieve comparable results. For SC 004, we need a complementary keyword. “Business” was an obvious choice, following the English component of the complex subject heading.

Tab. 2: Informatics publications in BASE

Search	Total SC 004	SC 004 + “business”	SC 330 + “Computer”	SC 650 + “computer”
# Results	1.325.373	31.278	12.693	29.039
Percent of BASE (314.978.975)	0,421 %	0,010 %	0,004 %	0,009 %

From this, we can conclude that publication activity in informatics is solid: 0,4 % of all papers in BASE represent a significant share. As in any other interdisciplinary platform, the largest share is typically with Medicine (DDC number 61 in BASE, equivalent for SC 610 in DNB), which is only four times larger.

Comparing the result set sizes for a post-coordinate search for business informatics, we find that 2,5 and 2,25 times more papers indexed under SC 004, respectively SC 650, compared with those in SC 330

Even more interesting was that 18 of the top 20 results (by relevance) were coming from German sources (ZBW and DNB) when searching for SC 330 + “computer”.

While these results are certainly not representative, they indicate that SC 330 is not a comprehensive category when looking for papers about business informatics. We also can say the international community seems to be doing something different than the German cataloguing rules would call for.

54 https://www.base-search.net/.

55 https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/DDC/ddcSachgruppenDNBAb2013.pdf.

4 Conclusions and outlook

Using thesauri as an indexing system provides reproducible and explainable results. Using quantitative and qualitative indicators provided by the machine, we can understand why a specific indexing result or also the indexing results for a group of documents was successful or not. Specifically relevant for this is the explicit context within the index but also within the indexate, i. e., we can understand in detail how a prediction or decision was taken by the machine.

The characteristics is specifically useful for data validation, identifying candidates for potential data repair. In many cases, we can identify a candidate automatically by simply comparing predicted and expected values in the indexate. A thesaurus or any other KOS following applicable librarian cataloging rules for the task will assist us in this matter with assertions based on the encoded rule set and this can be automated, using the proposed machine.

The main virtues leading to a high significance of this method as a reference include the application to the full text of the document. We should not forget that machines outperform human subject catalogers when it comes to a complete synopsis of the full text. Even if manual autopsy methods are effective and experience of a subject specialist cannot be made up by machines just yet, a full text indexate against the relevant subject headings' vocabulary is valuable auxiliary information that no responsible expert would want to ignore.

Sometimes, we may find the existing thesaurus may not answer all the questions. As with any other approach, some disciplines the GND subject headings are more elaborate than in others. This is a challenge we may share with conventional approaches. We can however confirm that quality is fair or good for most disciplines with relevant publication activity.

However, in some cases the method even uncovers some challenges with the rule set per se. While publication activity in informatics is substantial, the indexing vocabulary for that discipline and related disciplines is sparse and partly misaligned. Specifically, this seems to be a problem for business informatics in GND.

We have observed how sparse and incomplete coverage of the domain discipline's knowledge representation in the thesaurus results in bias. Additionally, we have established the notation for business informatics is inconsistent with the remaining vocabulary of corresponding domains, the existing DDC notation is probably not comprehensive enough and the approach for the discipline in GND is different from the international context.

As a result, authors in the area may be difficult to find in the territory governed by GND (German). Business informatics publications, as any other topic as well, can only be found if relevant search entries align with the hierarchy of the respective classification system.

The decisions made for classifying business informatics under Economy do not only affect publications in German, as the indexing rules are rather a cultural-regional tradition than language dependent.

Questions that yet must be answered, include:

– Under which subject category should we find publications about business informatics?
– Can we establish a representative corpus of publications under the current classification rules?
– How can we establish a sufficient access vocabulary for the discipline?

Measures that could help with a solution of the problem may include

– Validating the Dewey notations of the related vocabulary in GND,
– Corpus analysis of the representative publications,
– Integration of domain-specific vocabulary into GND (i. e., from STW),
– Comprehensive analysis of significant subject headings found in other languages and re-conciliation of the DDC notations for their keywords.

Eventually, also this work confirms that universal classification is still facing the challenge of sufficient coverage of universal knowledge. Regardless of the method, we sometimes may find ourselves restricted by incomplete models or bias coming from publication frequency distribution. This we can only overcome by well-balanced efforts for the models we need. The ongoing evolution of authority data and the related cataloguing rules are an incremental, yet infinite process and we must raise and maintain awareness among the stakeholders of the relevance of thesaurus work.

References

Gabler, Sebastian (2021): Vergabe von DDC-Sachgruppen mittels eines Schlagwort-Thesaurus. Universität Wien. Available at https://utheses.univie.ac.at/detail/60927#.

Gödert, Winfried; Hubrich, Jessica; Nagelschmidt, Matthias (2014): Semantic Knowledge Representation for Information Retrieval. Berlin: De Gruyter.

Golub, Korjalka; Hagelbäck, Johan; Ardö, Anders (2016): Potential and Challenges of Subject Access in Libraries Today on the Example of Swedish Libraries. In: *International Information & Library Review*, 48 (3), 204–10.

Golub, Korjalka (2019): Automatic subject indexing of text. In: *Encyclopedia of Knowledge Organization*, ed. by B. H. and C. Gnoli.

Available at https://www.isko.org/cyclo/automatic, accessed 2022-11-30.

de Jesus Holanda, Adriano; Torres Pisa, Ivan; Kinouchi, Osame; Souto Martinez, Alexandre; Seron Ruiz, Evandro Eduardo (2004): Thesaurus as a complex network. In: *Physica A: Statistical Mechanics and its Applications*, 344 (3–4), 530–36. DOI:10.1016/j.physa.2004.06.025.

ISO 25964-1 International Organization for Standardization (2011): Information and documentation – Thesauri and interoperability with other vocabularies –Part 1: Thesauri for information retrieval (ISO Standard No. ISO 25964-1:2011). Available at https://www.iso.org/standard/53657.html.

ISO 25964-1 International Organization for Standardization (2013): Information and documentation – Thesauri and interoperability with other vocabularies –Part 2: Part 2: Interoperability with other vocabularies (ISO Standard No. ISO 25964-2:2013). Available at https://www.iso.org/standard/53658.html.

DNB Deutsche Nationalbibliothek (2017): Regeln für die Schlagwortkatalogisierung. 4., vollständig überarbeitete Auflage. (RSWK) Available at http://nbn-resolving.de/urn:nbn:de:101-2017011305.

DNB Deutsche Nationalbibliothek (2019): Grundzüge und erste Schritte der künftigen inhaltlichen Erschließung von Publikationen in der Deutschen Nationalbibliothek. Available at https://www.dnb.de/SharedDocs/Downloads/DE/Professionell/Erschliessen/konzeptWeiterentwicklungInhaltserschliessung.html, accessed 2023-01-11.

Mödden, Elisabeth; Tomanek, Katrin (2012): Maschinelle Sachgruppenvergabe Für Netzpubikationen. In: *Dialog mit Bibliotheken*, 24 (1), 17–24. Frankfurt, Leipzig: Deutsche Nationalbibliothek.

Nentidis, A.; Krithara, A.; Katsimpras, G.; and Paliouras, G. (2022): Overview of BioASQ task a Large-Scale Online Biomedical Semantic Indexing. 6. Workshop Computerunterstützte Inhaltserschließung. Available at https://wiki.dnb.de/download/attachments/252121510/BioASQ10_overview_taska%20%281%29.pdf, accessed 2022-12-26.

Pellegrini, Tassilo; Blumauer, Andreas (2006): Semantic Web und semantische Technologien: Zentrale Begriffe und Unterscheidungen. In: *Semantic Web. Wege zur vernetzten Wissensgesellschaft*, ed. by Tassilo Pellegrini and Andreas Blumauer. Heidelberg: Springer.

Roget, Peter Mark (1856): Thesaurus of English Words and Phrases so as to facilitate the expression of ideas. London: Longman, Brown, Green, and Longmans.

Uhlmann, Sandro (2013): Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der GemeinsamenNormdatei (GND). In: *Dialog mit Bibliotheken*, 25 (2). Frankfurt, Leipzig: Deutsche Nationalbibliothek.

Sebastian Gabler, MSc.
Chief Customer Officer
Semantic Web Company GmbH
Hollandstrasse 15
A-1020 Wien
Österreich
sebastian.gabler@sebastian-gabler.eu
https://orcid.org/0000-0002-8607-0952