JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12775

JLIS.it

Artificial intelligence, machine learning and bibliographic control. DDC Short Numbers – Towards machine-based classifying

Elisabeth Mödden[a]

a) Deutsche Nationalbibliothek, http://orcid.org/0000-0001-6809-3926

Contact: Elisabeth Mödden, e.moedden@dnb.de

Received: 25 August 2021; **Accepted:** 20 September 2021; **First Published:** 15 January 2022

ABSTRACT

Digital publications now account for the majority of new accessions at the German National Library each year. Due to this growing number, it has become quite challenging to collect and catalogue these items properly. At the same time, these changes allow for new ways, in which it can use the collections. For a number of years, the DNB has been addressing the question of how subject cataloguing processes can be automated so that bibliographic records can be enriched with metadata as comprehensively and uniformly as possible. In the course of introducing automated subject cataloguing procedures, work is also being done on the automated assignment of Dewey Decimal Classification numbers. For this purpose, a set of abridged DDC numbers based on is being developed. The article sheds light on how artificial intelligence is used in this process. Furthermore, the challenges posed by the development of DDC short numbers and machine-based classification for different scientific subjects will be addressed. Also, it discusses how the DNB deals with the issues of data provenance, data delivery and quality management.

KEYWORDS

Dewey Decimal Classification; DDC Short numbers; Artificial intelligence; Machine-based classification.

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12775

Introduction

In the German National Library (Deutsche Nationalbibliothek / DNB), both verbal and classificatory subject cataloguing are used for subject indexing. In the course of introducing automated subject cataloguing procedures, work is also being done on the automated assignment of Dewey Decimal Classification numbers. For this purpose, a set of abridged DDC numbers based on, but not limited to, the DDC Abridged Edition 15 and hereafter referred to as DDC Short Numbers, is being developed.

First experiences in the automatic assignment of abridged numbers were gained in the field of medicine (DDC 610). Since 2005, medical dissertations have been classified using a set of 140 DDC Short Numbers. Since 2015, these Short Numbers have been assigned automatically by utilizing artificial intelligence. Short Number sets for other DDC areas are currently being developed. It is planned to extend the automatic assignment of Short Numbers to all subjects and to constantly review the process and its results.

Initial situation

Digital publications now account for the majority of new accessions at the German National Library each year, and the number is rising (see figure 1). In 2020, the collections grew by approx. 1 million online publications like e-books and electronic journal articles. Due to this growing number, it has become quite challenging to collect and catalogue these items properly. At the same time, these changes allow for new ways, in which we can use our collections; for example, it is possible to search for and retrieve individual articles.

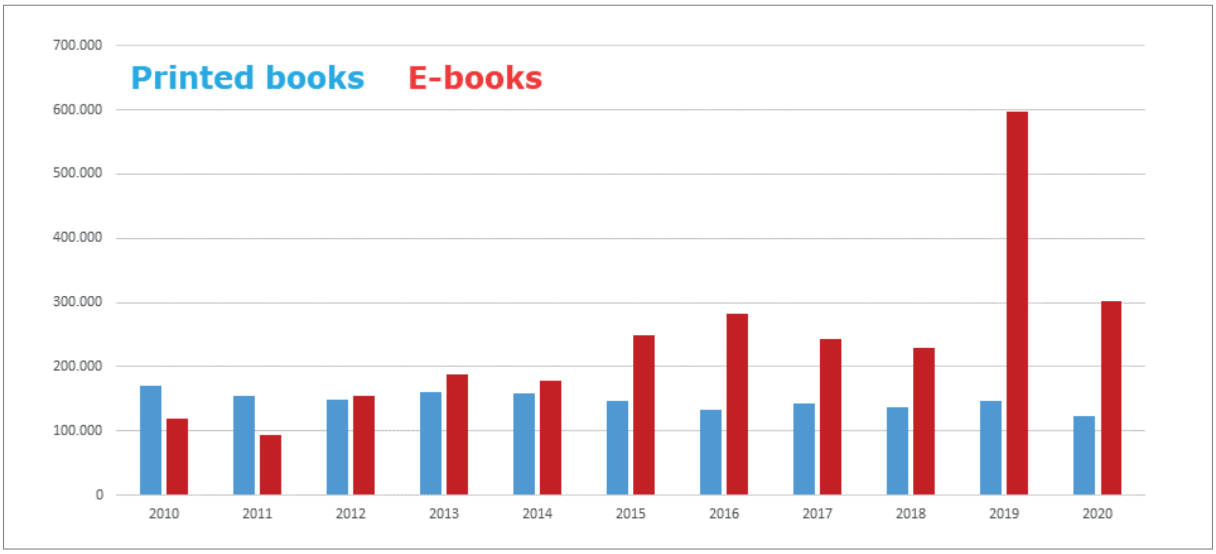


Fig. 1. Increasing amount of publications to be catalogued (e. g. monographs)

Subject cataloguing makes it possible to structure the library's large collections thematically and thereby facilitate the retrieval of publications in these collections. For a number of years, the

DNB has been addressing the question of how subject cataloguing processes can be automated so that bibliographic records can be enriched with metadata as comprehensively and uniformly as possible – despite new media formats and ever-increasing number of units. Other advantages of automated processes, e.g. the possibility of cataloguing component part of works such as the above-mentioned journal articles both by classification and by assigning subject headings, should be exploited consistently.

Since 2010, the DNB has increasingly been classifying and indexing digital publications using automated procedures rather than intellectual processes (Gömpel, Junger, and Niggemann 2010). In September 2017, the use of machine-based cataloguing procedures was extended to physical publications (Junger and Schwens 2017) ("Cataloguing Media Works" n.d.). In the DNB's Strategic Compass 2025 (Deutsche Nationalbibliothek 2016a) and Strategic Priorities (Deutsche Nationalbibliothek 2016b), the reorganisation of subject cataloguing is addressed as a significant area of activity that will continue to be important during the years to come. This article sheds light on how artificial intelligence is used in this process. Furthermore, the challenges posed by the development of DDC short numbers and machine-based classification for different scientific subjects will be addressed. Also, it discusses how the DNB deals with the issues of data provenance, data delivery and quality management.

Cataloguing methods

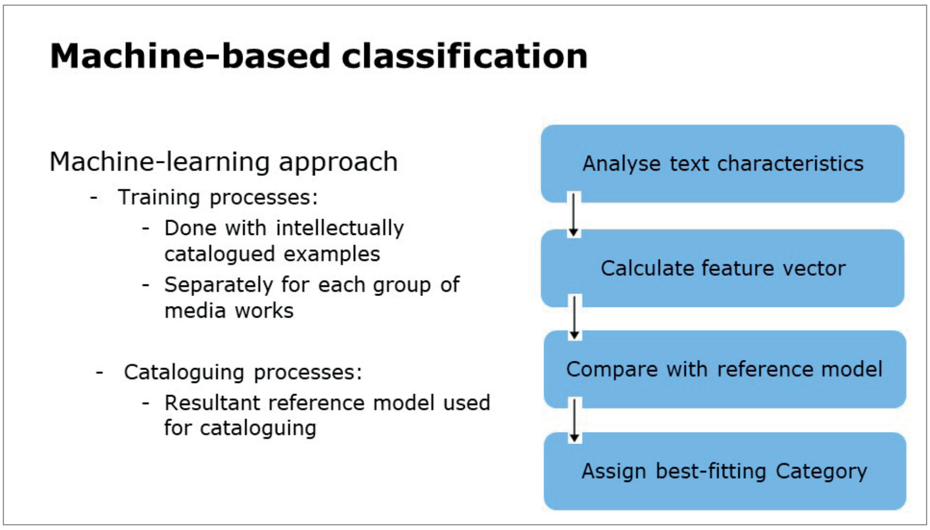
Subject cataloguing at the DNB is based on the Series of the Deutsche Nationalbibliografie (German National Bibliography). Every publication catalogued since the bibliographic year 2004 is assigned to one of roughly one hundred subject categories, which are organised in accordance with the Dewey Decimal Classification (DDC) system ("Dewey Decimal Classification (DDC)" n.d.). Beyond that, the publications from the publishers´ book trade provided in Series A are processed intellectually using built numbers from the DDC and subject headings from the Integrated Authority File, the Gemeinsame Normdatei ("Gemeinsame Normdatei (GND)" n.d.).

The development of software applications for subject cataloguing purposes started with the PETRUS project (Schöning-Walter 2010). Machine-based subject category assignment began in 2012, while the automated assignment of subject headings got under way in 2014. Medical publications were first automatically assigned DDC Short Numbers in 2015. At present, work is under way to develop DDC Short Numbers for all subjects.

The DNB employs a support-vector machine for the use in machine-learning processes to facilitate automated classification using DDC Subject Categories and DDC Short Numbers (Mödden and Tomanek 2012). The characteristics of selected text parts and existing metadata are analysed by means of linguistic and statistical methods. During the training phase, the system analyses publications with intellectually assigned Subject Categories and Short Numbers to generate a reference model for all classes of DDC short numbers. When creating this model, it is essential that each class contain sufficient numbers of appropriate learning examples. During the cataloguing process, the system then calculates a statistical measure to determine how closely the content of a new publication matches the patterns learned. As the result of topical classification, the best-matching Subject Categories and Short Numbers are assigned to the publication (see figure 2).

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12775

The cataloguing software was created in cooperation with the Freiburg-based company Averbis and is integrated into the DNB's system infrastructure. Machine-based classification has been implemented for texts in German and English.



Machine-based classification

Machine-learning approach

- Training processes:
  - Done with intellectually catalogued examples
  - Separately for each group of media works
- Cataloguing processes:
  - Resultant reference model used for cataloguing

Analyse text characteristics

Calculate feature vector

Compare with reference model

Assign best-fitting Category

Fig. 2. Processes used in machine-based classification

Workflow

In productive operations, the machine-based cataloguing process (see figure 3) begins automatically at a fixed time every day by sending a list of publications that require first-time processing [1] to a web service. This service retrieves the existing metadata [2] from the cataloguing database (CBS) and the digital full text files or tables of contents [3] from the repository. Before being transmitted to the cataloguing software [4], the storage formats are converted into simple text files and the main language of the publication is determined. Once they have been processed in this way, the results of the analytical process [5] are added to the publication's bibliographic record [6]. Anomalies found during processing are recorded in system files.

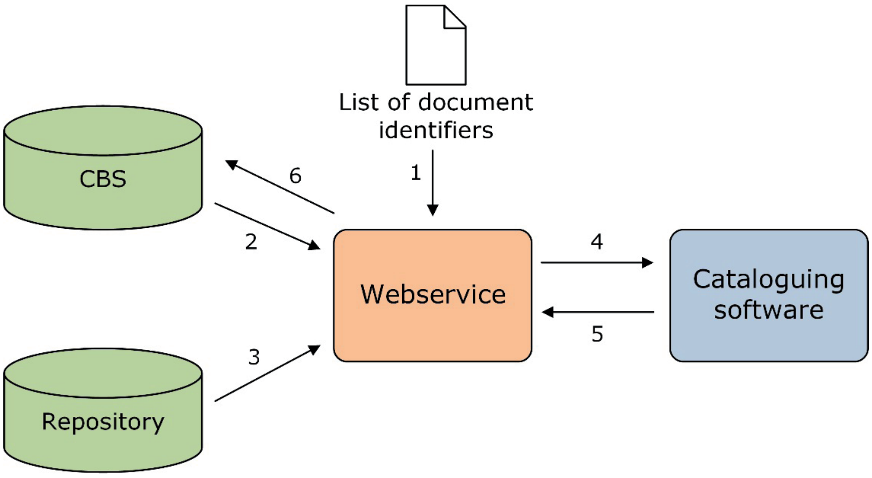


Fig. 3. Technical process used at the DNB for automated cataloguing (productive operation)

JLIS.it vol. 13, no. 1 (January 2022)
ISSN: 2038-1026 online
Open access article licensed under CC-BY
DOI: 10.4403/jlis.it-12775

The cataloguing software has various configuration options enabling different types of publications to be processed in different ways. These configurations consist of parameter settings, which have been optimised during test runs. They facilitate the identification of the classification model, for example when assigning Subject Categories. Depending on the features of the publication a certain configuration is set: for instance, digital monographs are processed differently from journal articles, German-language texts are processed differently from English ones, full text files are processed differently from digitised tables of contents.

The software, training corpora and added GND vocabulary undergo regular maintenance and development to ensure that the system as a whole improves constantly. At certain times, digital publications catalogued intellectually are added into the machine-learning processes as new examples. This raises the question whether automated cataloguing processes should be repeated when significant progress is made. In the future, we want to introduce cyclical repetition in order to improve the quality of our machine-generated metadata. We also want to include publication formats that previously were omitted.

Milestones

At the beginning of 2017, the automated cataloguing processes were extended to include journal articles in digital format. The import procedure for e-journals started at the beginning of 2016. Around 675,000 journal articles were integrated into the DNB's collections in 2016 alone. Beginning with journals published by Springer Publishing, the DNB is now enriching individual articles with subject cataloguing metadata. In view of the great number, periodical online publications can only be catalogued economically at this extent by using automated methods.

Another strategic milestone was reached in September 2017 when the automated cataloguing processes were extended to printed monographs in the Deutsche Nationalbibliografie's H Series ("Deutsche Nationalbibliografie" 2019)[1]. Since September 2017 the DNB no longer applies full DDC numbers for this Series. These are to be gradually replaced by DDC Short Numbers. Publications from the publishers' book trade (Series A) will continue to be catalogued intellectually.

In due time, all existing digital resources, for example parallel online editions, tables of contents, abstracts, blurbs and cover texts, will be used for machine-based subject cataloguing of physical media works. At present, publications are catalogued on the basis of digitised tables of contents and the bibliographic metadata that has been supplied. However, since there is less text and a lack of substantial information in some tables of contents the conditions for text analysis are frequently more unfavourable than in the case of online publications. Therefore, the automatically assigned Subject Categories for Series H are all reviewed intellectually.

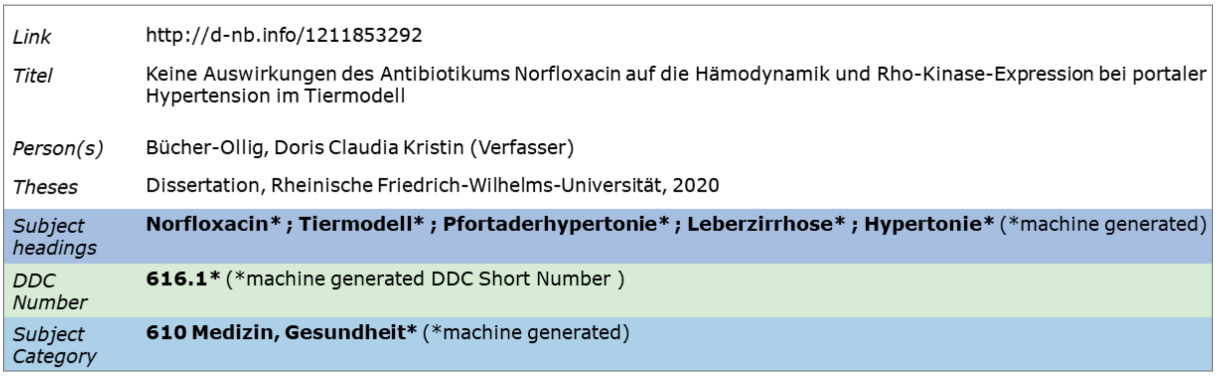
[1] In Series H are university publications: Dissertations and postdoctoral theses from German universities and German-language dissertations and postdoctoral theses from abroad.

How DDC Short Numbers are selected

Until they are ready for productive use in automatic classificatory indexing, DDC Short Numbers have to pass a multi-stage workflow. Dewey numbers are selected per subject, using the DDC Subject Categories as a guide. This process is accompanied, if necessary, by a comparison with Dewey numbers of DDC's Abridged Edition 15. The next step is to analyse the frequency of occurrence of DDC numbers on the basis of the literature published over the last ten years. Building on this, suitable numbers are selected while numbers with low literature warrant are discarded. This data set is then used for initial technical tests to see how well the selected numbers are working for automatic assignment in the respective subject. Mismatches are analysed and Short Numbers are adjusted in an iterative process. Finally, if the results are convincing, the Short Numbers are put into productive operation and the selection process begins for the next Subject Category. The experts at the Department for Subject Indexing are closely monitoring this iterative process.

Provenance data

The decision to apply automatic processes goes along with the decision to assign the machine-generated metadata to the bibliographic record, to display it in the DNB portal, to use it for retrieval purposes, and to deliver it via the data services. In addition to this, metadata for journal articles is now available in the DNB catalogue and can be obtained through the data services. The DNB's database structure was modified to supply information on the provenance and reliability of the machine-generated metadata. In our database, the machine-generated metadata is recorded together with the date, the configuration name and the confidence value, which is an estimate of the data quality. The machine-generated DDC Short Numbers and subject headings are indicated as such when displayed in the DNB portal (see figure 4).



Link	http://d-nb.info/1211853292
Titel	Keine Auswirkungen des Antibiotikums Norfloxacin auf die Hämodynamik und Rho-Kinase-Expression bei portaler Hypertension im Tiermodell
Person(s)	Bücher-Ollig, Doris Claudia Kristin (Verfasser)
Theses	Dissertation, Rheinische Friedrich-Wilhelms-Universität, 2020
Subject headings	**Norfloxacin* ; Tiermodell* ; Pfortaderhypertonie* ; Leberzirrhose* ; Hypertonie*** (*machine generated)
DDC Number	**616.1*** (*machine generated DDC Short Number)
Subject Category	**610 Medizin, Gesundheit*** (*machine generated)

Fig. 4. Title of an automatically catalogued Series O publication displayed in the DNB catalogue with subject headings, DDC Short Number and DDC Subject Category

The data exchange format MARC 21 has also been modified so that standardised information on the provenance of the metadata can be distributed as well.

Quality and monitoring

Along with daily controls of the process operation, technical checks are carried out by means of sampling. Here, a selection of the publications submitted for automated cataloguing is also classified and assigned subject headings on an intellectual basis. All metadata generated during the cataloguing processes is recorded in the bibliographic database. For display and use in the portal and data services, preference is given to metadata assigned intellectually if available.

For quality management purposes, the quality of the machine-generated classifications is evaluated statistically by comparing automatically and intellectually assigned metadata. Existing metadata for parallel editions is also used for this purpose if applicable. Over the last five years, the DNB has reviewed approximately 18% of the automatically classified online publications in Series O ("Deutsche Nationalbibliografie" 2019). The machine-generated Subject Categories agreed with the intellectually assigned Subject Categories in 76% of cases. This average was actually clearly exceeded in some subject areas, e.g. in law (92% consistency) and medicine (87% consistency). However, machine-based classification does not yet function satisfactorily particularly in the case of subjects on which there is little literary warrant, because there is not enough of the training material required for the learning processes. One such subject for example is the history of South America (DDC Subject Category 980).

There are several issues with machine-based classification of DDC Subject Categories. The main problem is that the machine-assigned DDC Subject Category determines the Short Number. If the Subject Category is wrong, the Short Number will be wrong. Another challenge is posed by the fact that the DDC is continuously updated; even if changes on the broader hierarchy levels do not occur frequently, both changes in the meaning of the class (e.g. change of caption, added or removed major topics) and notational changes such as new or deleted numbers can have an impact on the correct assignment of a Short Number and thus must be taken into account in the process.

Combining machine-based and intellectual cataloguing

Automatic cataloguing procedures are not free from error. Along with imprecise or incorrect assignments, they also generate a bulk of metadata that is not useful for our patrons. The task of quality management is to critically evaluate the error ratio and its effects on the metadata stock in order to adjust the cataloguing processes if necessary. The goal is to achieve a high degree of reliability for the cataloguing data, irrespective of whether it was generated intellectually or automatically. The intellectual and machine-based processes are to be linked more closely in the future. Quality management serves to control and determine which publication forms can be catalogued automatically and which cataloguing services have to be performed intellectually.

Outlook

In 2018, the company Averbis announced a stop to further software developing for machine-based cataloguing. The existing software will be supported only for the next 5 years. Thus, the development of a new machine-based cataloguing system is under way. The target is a new software with a modular structure. This will make it easier to replace individual tools in the future. For this purpose, the project "Erschließungsmaschine" – EMa was started. By this, the Averbis software is scheduled to be replaced with a new modular software system by 2022.

Major requirements for the new system are individual modules for text extraction, language recognition, classification, subject indexing, management of text corpora, of terminologies and of notations, etc. After a detailed market study, the Annif toolkit was selected. The National Library of Finland has developed Annif as a tool for machine indexing. The open-source toolkit "uses a combination of existing natural language processing and machine learning tools including Maui, Omikuji, fastText and Gensim. It is multilingual and can support any subject vocabulary (in SKOS or a simple TSV format). It provides a command-line interface, a simple Web UI and a microservice-style REST API." ("Annif – Tool for Automated Subject Indexing" n.d.). For more details, see the very interesting paper by Osma Suominen (Suominen 2019) and the Documentation on GitHub ("GitHub – NatLibFi/Annif:." n.d.). The DNB is very much looking forward to working with Annif, since it is a very promising new tool and is firmly believing that it will pose new opportunities for machine-based classifying and indexing.

In addition, a new, innovative AI (artificial intelligence) project is being launched at DNB. The DNB wants to develop new methods for processing and analysing content and metadata. The new approach should improve the quality of machine-based content indexing in a significant way. Potential AI developments, which are suitable for cataloguing text-based publications, will be investigated, selected, combined and adapted. Research will be conducted to determine which AI methods can be used for machine processing and analysis of natural language texts in order to obtain the most complete and accurate indexing data. The DNB aims for flexibly reusable tools (open-source tools), so that other libraries or institutions with comparable tasks can use these developments as well.

A good database, based on high-quality intellectual indexing by subject experts, is an indispensable prerequisite for the AI project. Therefore, the Department for Subject Indexing will be intensively involved in the development of new procedures. Furthermore, the rules for subject cataloguing should be adapted in such a way as to benefit the combination of both approaches – intellectual and machine-based subject indexing. In the end, the DNB is convinced that high-quality indexing can be achieved by combining intellectual and machine generated classifying and indexing.

References

"Annif – Tool for Automated Subject Indexing". n.d. Accessed 30 July 2021. http://annif.org/.

"Cataloguing Media Works". n.d. Accessed 29 July 2021. https://www.dnb.de/EN/Professionell/Erschliessen/erschliessen_node.html.

"Deutsche Nationalbibliografie". 2019. https://www.dnb.de/EN/Professionell/Metadatendienste/Metadaten/Nationalbibliografie/nationalbibliografie.html.

Deutsche Nationalbibliothek. 2016a. *2025: Strategic Compass*. Leipzig, Frankfurt, M: Deutsche Nationalbibliothek. https://d-nb.info/1112299556/34.

Deutsche Nationalbibliothek. 2016b. *Strategic Priorities 2017–2020*. Leipzig, Frankfurt, M: Deutsche Nationalbibliothek. https://d-nb.info/1126595101/34.

"Dewey Decimal Classification (DDC)". n.d. December. Accessed 30 July 2021. https://www.dnb.de/EN/Professionell/DDC-Deutsch/ddc-deutsch_node.html.

"DNB_Strategic-Compass-2025_lesesprache_englisch.Pdf". n.d.

"Gemeinsame Normdatei (GND)". n.d. Deutsche Nationalbibliothek. Accessed 30 July 2021. https://www.dnb.de/DE/Professionell/Standardisierung/GND/gnd_node.html.

"GitHub - NatLibFi/Annif: Annif Is a Multi-Algorithm Automated Subject Indexing Tool for Libraries, Archives and Museums. This Repository Is Used for Developing a Production Version of the System, Based on Ideas from the Initial Prototype." n.d. GitHub. Accessed 30 July 2021. https://github.com/NatLibFi/Annif.

Gömpel, Renate, Ulrike Junger, and Elisabeth Niggemann. 2010. "Veränderungen Im Erschließungskonzept Der Deutschen Nationalbibliothek". *Dialog Mit Bibliotheken* 22 (1): 20–22.

Junger, Ulrike, and Ute Schwens. 2017. "Die Inhaltliche Erschließung Des Schriftlichen Kulturellen Erbes Auf Dem Weg in Die Zukunft". *Dialog Mit Bibliotheken* 29 (2): 4–7.

Mödden, Elisabeth, and Katrin Tomanek. 2012. "Maschinelle Sachgruppenvergabe Für Netzpublikationen". *Dialog Mit Bibliotheken* 24 (1): 17–24.

Schöning-Walter, Christa. 2010. "PETRUS – Prozessunterstützende Software Für Die Digitale Deutsche Nationalbibliothek". *Dialog Mit Bibliotheken* 22 (1): 15–19.

Suominen, Osma. 2019. "DIY Automated Subject Indexing Using Multiple Algorithms". *LIBER Quarterly* 29 (1): 1–25. https://doi.org/10.18352/lq.10285.

"The Integrated Authority File (GND)". n.d. December. Accessed 29 July 2021. https://www.dnb.de/EN/Professionell/Standardisierung/GND/gnd_node.html.