IFLA WLIC 2018
KUALA LUMPUR

Submitted on: 19.07.2018

Automation first – the subject cataloguing policy of the Deutsche Nationalbibliothek

Ulrike Junger
Head Domain Acquisition and Cataloguing, Deutsche Nationalbibliothek, Frankfurt/Main, Germany.
E-mail address: u.junger@dnb.de

Abstract:

In 2006, the legal mandate of the German National Library (Deutsche Nationalbibliothek/DNB) was expanded to digital publications. It soon became clear that the traditional cataloguing process could not be applied to deal with the vast amount of digital publications collected. In 2010 it was therefore decided to re-use metadata provided by the publishers to create basic catalogue records. Since subject data are often not provided, the DNB also in 2010 started a project to develop methods of automated assignment of subject headings and DDC-based subject categories. These automated procedures are now routinely used to assign subject access points to digital publications. In 2017 the DNB took the next step and devised a new policy for classifying and indexing all of its collections. The principal goal is to create subject access points using automated procedures as principal method, also for printed material. Intellectual subject cataloguing should in the future only be done when automated procedures yield no or qualitatively insufficient results. In this context DNB also decided to discontinue the regular application of the Dewey Decimal Classification including number building. Instead, a system of short DDC notations is designed with the intention to automated assignment.
The current state of the implementation of this policy and the results so far achieved are described as well as the methods and outcomes of the automated procedures. The challenges coming along with this approach include the necessity of developing new tools for updating the subject authority file, quality issues and a transformation of workflows.

Keywords: subject cataloguing; automatic indexing; automatic classification

Introduction

The Deutsche Nationalbibliothek (DNB) is the central archival library and national

bibliographic center of the Federal Republic of Germany. Its tasks and function are laid out in the Law regarding the Deutsche Nationalbibliothek: collecting, permanently archiving, bibliographically indexing and making the collections available to the general public. The DNB's legal collection mandate encompasses all publications in text, image and sound which were published in Germany or in German language from 1913 onwards, also foreign publications about Germany, translations of German works, and the works of German-speaking emigrants published abroad between 1933 and 1945. Also, to collect are sheet music and sound recordings. In the year 2006, the legal mandate of the DNB was expanded to digital publications. Especially after successfully developing and implementing procedures for bulk ingest of digital publications it became clear that the traditional cataloguing process could not be applied to deal with the vast amount of digital publications collected. Figure 1 shows the development of the collection of printed and online monograph indicating that although there is a slight decrease in the number of printed books in the recent years this is by far outweighed by the increase for digital books.

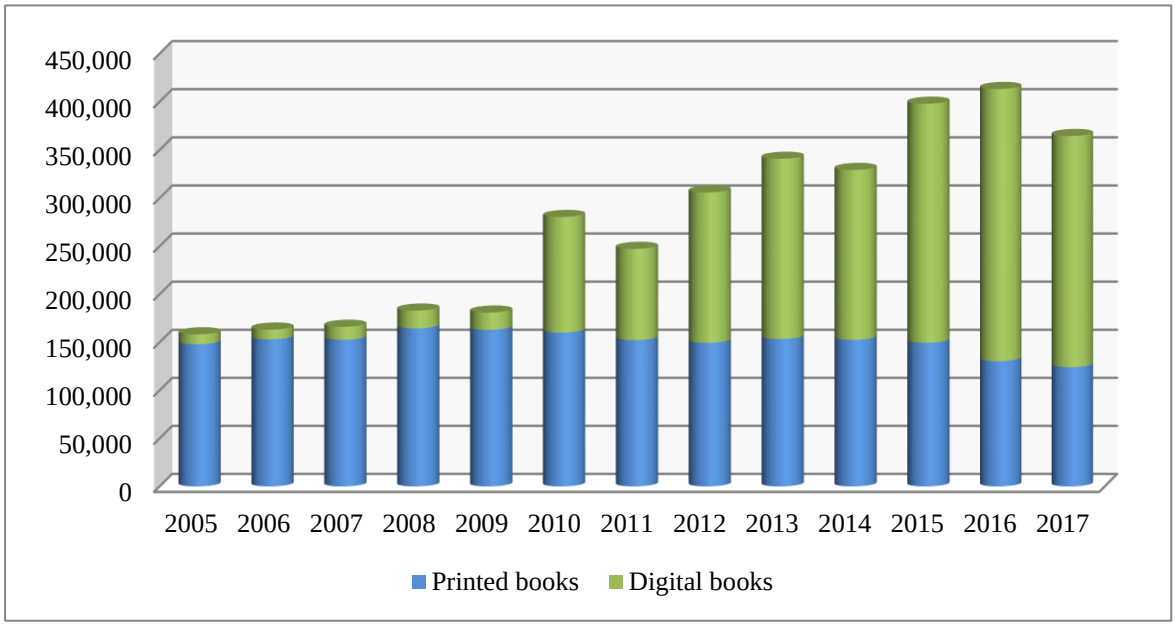


Fig. 1: Number of printed and online monographs collected in the DNB

In order to further fulfil the legal obligation regarding the national bibliography to 2010 it was therefore decided to re-use metadata provided by the publishers and depositors to create catalogue records. DNB requires depositors to deliver the digital object together with metadata using specific metadata core sets.[1]

Subject data are often not provided. Therefore, also in 2010 DNB started a project to develop methods of automated assignment of subject headings taken from the Gemeinsame Normdatei (GND/Integrated Authority File)[2] and subject categories[3] based on the Dewey Decimal Classification (DDC). Meanwhile, these automated procedures are successfully and routinely used to assign subject access points to online publications. In 2017 the DNB took the next step and devised a new concept for classifying and indexing[4]: The principal goal is to create subject access points for thematic retrieval purposes for all of its collections by

[1] http://www.dnb.de/EN/Netzpublikationen/Ablieferung/Metadaten/metadaten_node.html
[2] See http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html for further details.
[3] http://www.dnb.de/EN/Erwerbung/Inhaltserschliessung/sachgruppenDnb.html.
[4] http://www.dnb.de/EN/Erwerbung/Inhaltserschliessung/grundzuegeInhaltserschliessungMai2017.html.

principally using automated procedures for the creation of subject metadata. This is strived for also for printed books etc., based on digitized tables of contents and by re-using subject data created for digital parallel editions. Intellectual subject cataloguing should only be done when automated procedures yield no or qualitatively insufficient results, are required in conjunction with the management of the automated procedures or the authority file. This concept is now implemented step by step.

A basic principle of the subject cataloguing policy of DNB is, that the same systems should be used for both intellectual and automated subject cataloguing, i.e. the same set of subject headings and the same classificatory instruments. Regarding classification, it seems obvious that automated procedures cannot yield built DDC numbers. When the DNB decided last year to expand automated creation of subject metadata to printed material, it also decided to discontinue the regular application of the DDC including number building. Instead, a system of short DDC notations is designed which can be used both for intellectual and automated classification.

Current implementation status of DNB's subject cataloguing policy

Traditionally, the DNB has carried out a tiered model of subject cataloguing, not treating all publications alike. The assignment of a publication to a series of the Deutsche Nationalbibliografie determines its degree of being subject catalogued. Currently, the DNB is using both various methods (classification and subject indexing) and different processes (manual cataloguing by expert staff, automatic cataloguing and the re-use of data from other providers, i.e. third-party subject headings and notations which are integrated into the catalog record of a publication and also made available to users in the online catalog).

Classification includes the assignment of subject categories used to structure the Deutsche Nationalbibliografie (German National Bibliography) and of Dewey Decimal Classification (DDC) notations. Subject indexing is the assignment of subject headings and genre terms[5]. The subject headings are part of the Integrated Authority File (Gemeinsame Normdatei/GND). With the exception of topical headings, the authority records for person, corporate bodies, conferences and places are created according to Resource Description and Access (RDA). Topical terms are created using a specific rule-set called Regeln für die Schlagwortkatalogisierung (Headword Rules for Subject Cataloguing).

The current practice is as follows:

- For Series A (printed publications from the publishers' book trade) each publication is intellectually given a DDC subject category. Most publications also contain full DDC numbers[6].
- Series A as well as maps listed in Series C are also intellectually indexed with subject headings from the Integrated Authority File (GND).
- Publications in Series B (printed publications from outside the publishers' book trade) and H (printed dissertations) are catalogued using automated procedures based on scanned tables of contents. DDC subject categories are issued automatically or intellectually, subject headings automatically.
- Subject cataloguing of online publications (Series O) is only done automatically. With the exception of fiction and poetry and publications in other languages than German and

[5] http://www.dnb.de/EN/Erwerbung/Inhaltserschliessung/gattungsbegriffe.html.

[6] Not included here are school textbooks, fiction and poetry, children's and juvenile literature and medical dissertations.

English[7] all online publications receive a DDC subject category and for some subjects also DDC short notations. Subject headings are assigned to monographs and journals articles.

Methods of automated subject cataloguing

For automated subject cataloguing the DNB is operating the software system of a German company specializing in text mining and machine learning. It has been adapted to DNB's purposes and can be configured by the DNB to different groups of publications. In the following the some more details are given for the automated classification and indexing.

Automated Classification

The software used for classification is based on statistical learning algorithms, using a support vector machine. This is a mathematical process for pattern detection and recognition.

First a classification model has to be trained using intellectually classified publications. The DNB uses online publications, but also digitized tables of contents of printed books which make about two thirds of the training material. Currently only publications in German or English are used for training purposes.

After a linguistic pre-processing of the text features like specific vocabulary are extracted and related to the subject categories given to the documents. These patterns are generalized into a model. When the training is completed and satisfying results are achieved, the model is used for the classification of new publications. Classification might also be done retrospectively, i.e. for publications collected earlier but not automatically classified or for publications already classified but for which a re-classification promises better results, e.g. because of software improvements.

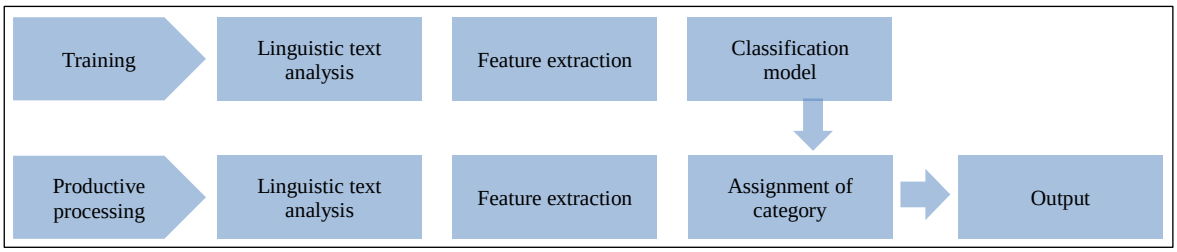


Fig. 2: Modelling for automated classification with statistical learning algorithms

Quality control is done in a two-fold manner: either by comparing the automatically assigned notation with an intellectually created notation taken from a parallel print edition of a publication or by the intellectual control of samples of automatically processed publications. The so-called F-measure[8] is calculated as a measure of quality.

When doing quality control, incorrect notations are corrected and the results of this control feed back into the training of the classification models. Sources of errors are too few documents available for training, nonspecific vocabulary in publications and the appearance of new subjects and concepts not reflected in the classification.

There are currently two implementation scenarios for automated classification in the DNB. The automated assignment of DDC subject categories for online publications has been done

[7] These account for about 35 % of the publications collected.
[8] The F-measure is the harmonic average of recall and precision.

since 2012. Processed are monographic online publications, journal articles and - since September 2017 - also printed dissertations and publications from outside the publishers' booktrade with digitized tables of contents. In all groups both German and English publications are processed.

Currently, between 5.000 to 10.000 publications are categorized per month, with an average F-measure of 0.7 – 0.8. Altogether so far more than 1.2 million publications were categorized.

The second and more ambitious scenario is the assignment of DDC short notations. Since the DNB started using the DDC in 2007, short notations have been used classifying medical dissertations. A set of about 140 notations was designed to cut work on classifying the around 10.000 medical dissertations the DNB collects annually. In 2015 the DNB started experimenting with the automated assignment of these medical short notations to online publications with promising results. This as well as the aim to also apply automated procedures to physical publications lead – as mentioned - to the decision to devise DDC short notations for all areas. The goal is to classify all materials collected by DNB in a homogenous way that is more specific than subject categories.

The design of the DDC short notations is executed by DNB's subject specialists. Based on the DDC Abridged Edition 15 notations are selected, reflecting the literary warrant of the DNB in the last 10 years. In a repeated process, adaptations are done, eliminating notations not needed and adding further notations. All short notations are compatible with the DDC. The number and length of notations varies from subject to subject, also depending on what could be found in the DDC. The set of short notations in chemistry comprises 13 classes, whereas social sciences has 92 classes.

DDC subject category	300 (sociology)
Short notations	303.6 (conflict and conflict solving)
	306.82 (mating patterns)

DDC subject category	610 (medicine)
Short notations	615.532 (homeopathy)
	618.92 (pediatrics)

Fig. 3: Examples for DDC short notations

Currently the automated assignment of DDC short notations is done for online publications for the subject categories medicine, chemistry, computer science and sociology, for medical dissertations and medical grey publications also for printed material.

Altogether by the end of 2017 around 260.000 online publications were classified with DDC short notations, most of them medical publications, and about 2.500 print publications.

Results show that for medical monographs an F-measure of 0,7367 resulted, based on samples comprising 11.808 titles. For computer science an F-measure of 0,8817 was calculated, based on a much smaller sample of 48 titles, taken from a pool of 1.135 publications.

It is intended to automatically assign short notations for more subjects successively, whenever the design of the short notations and their testing are completed. The switch from full DDC notations to short notations in intellectual classification will be done all at once.

Automated Indexing

In contrast to classification the method for automated assignment of subject headings to publications is not based on machine learning, but on computer linguistic algorithms and the use of a dictionary, which is part of the processing software. The dictionary contains over 1 million terms and is composed of semantic concepts taken from the GND (topical terms, persons, places, corporate bodies, conferences, names of works).

The indexing process involves several steps as shown in figure 4. Basis of the linguistic analysis are bibliographical title data and the available digital text (full text or digitized tables of contents). Various analysis methods like sentence detection, part-of-speech tagging, chunking etc. are applied to retrieve concepts from the text conveying its content. The result is a list of terms which are possible candidates for headings. They are matched against the concepts in the dictionary. A weighted selection of the candidate terms eventually results in a list of up to 7 subject headings. The bibliographical record is then linked to the authority records of these headings. The authority file which contains multiple inter-record links can thus be used for subject access of publications indexed automatically. Users can use the same vocabulary for the search of both printed and online material.

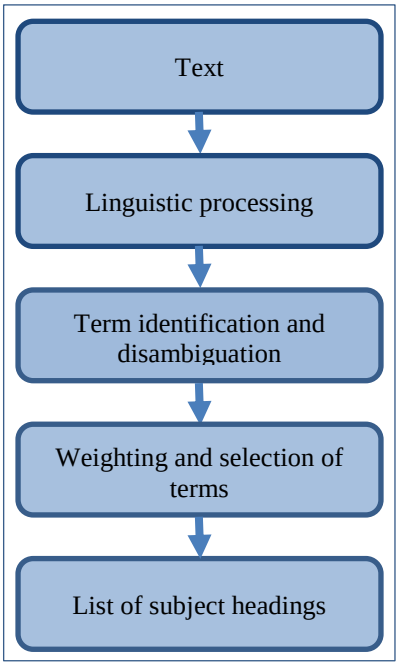


Fig. 4: Process steps of automated indexing

A major challenge in automated indexing is the dealing with ambiguous terms. The DNB has developed a sophisticated set of measures for disambiguation of homonyms. An important factor in that respect is that all subject headings in the GND (with the exception of place names) are categorized. This classificatory information is related to the subject categories assigned to the publication to be indexed.

An example: there are several authority records for persons with the name Max Weber: a sociologist, an actor, a politician, an artist. If the name “Max Weber” occurs in a text, it is likely that the record for the sociologist is correct if the content of the publication was classified as belonging to social sciences.

The DNB applies several modes of quality control with respect to automated subject indexing:

- The indexing process is continuously monitored by staff looking for systematic errors and failed recognition of terms. Errors are categorized and may lead to changes in the dictionary (e.g. terms can be de-activated or missing terms added).
- Samples of headings are checked for consistency of assignment.
- Samples of parallel editions are compared, with one edition (usually the print one) indexed intellectually and one edition (the digital one) indexed automatically.
- DNB's subject specialists do intellectual quality control for samples of publications indexed automatically. Headings are rated with respect to usefulness for searching and retrieving the title or a subject. The scale is 4-fold, reaching from very useful, useful, less useful to wrong/misleading. Figure 5 shows the results for online publications of the year 2016 indexed automatically.

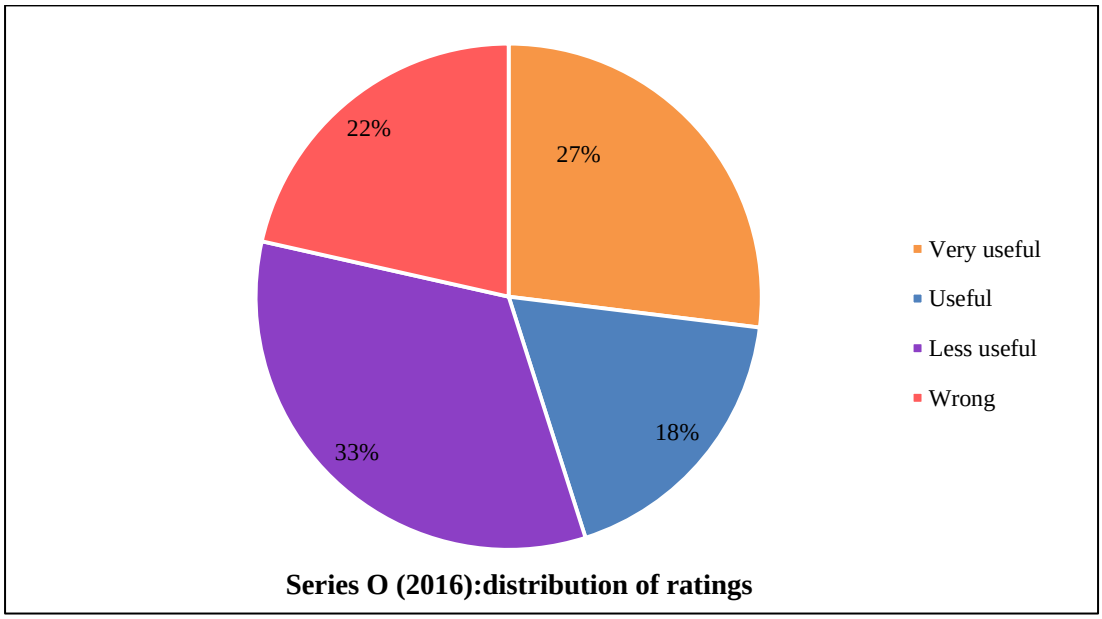


Fig. 5: Ratings for usefulness of automatically generated subject headings for online publications 2016

Reasons for indexing faults are multiple. Disambiguation may have failed or there is no suitable heading on the authority file, so a wrong heading was selected. The linguistic analysis resulted in incorrect indexing or a heading is correct but irrelevant for the contents of the publication. Measures for improvement are multiple as well: Curation of the dictionary, i.e. by setting a term on "ignore", altering parameters in the configuration of the indexing software, e.g. the number of headings to be issued, enrichment of the vocabulary used for indexing and of course improvements in the process itself, e.g. better disambiguation mechanisms.

Between 2014, when the DNB started with the automated indexing of online dissertations and scientific online monographs and the end 0f 2017, about 182.000 publications were automatically enriched with subject headings, thereof about 153.500 online monographs, 21.500 digital journal articles and around 7.000 printed monographs.

So far only publications in German language can be indexed. The DNB is therefore conducting a project to index English language material which makes up a considerable

portion of the collection with English language subject headings. The vocabulary chosen as indexing language are the topical headings from the Library of Congress Subject Headings (LCSH) which were loaded into the dictionary of the indexing software. Names of persons, corporate names and conferences are taken form the GND. So far, this process has not gone into regular production because more testing is needed.

Issues and challenges

Editing and management of the authority file

Traditionally, the German subject headings authority file is edited depending on the literary warrant, i.e. a heading is introduced, enriched or altered by subject catalogers when needed for indexing a specific document. Finding ways to edit and manage the authority file is also a necessity with automated indexing for only an up-to-date vocabulary will allow adequate indexing. The DNB is therefore currently conducting a project to build a so-called suggestion system: the goal is that during the automated indexing process possible candidates for subject headings are identified and presented to the authority file editors for editing.

Quality discussion

DNB's new subject cataloguing policy set off a discussion about the requirements regarding quality of subject data with respect to modern information retrieval systems. How is quality regarding subject cataloguing defined and what are proper criteria and instruments for measurement? The DNB is discussing these questions with its partners in the German-speaking library community.

Besides, as with the automated procedures themselves, the library is working on the continuous improvement of workflows and tools for quality management. Also it is intended to commission retrieval tests in order to assess the effects of less useful or incorrect rated subject headings and notations on thematic searches.

Changes of workflows

Introducing automated procedures for metadata generation implies also a change in work routines. Sample-based quality control has already become a new permanent task for subject cataloguers at the DNB. Whereas traditional intellectual cataloguing means the individual treatment of an individual copy of a work, automated procedures require also dealing with data sets and doing quality assessment both for individual documents but also systematically. Additional skills and knowledge are required, e.g. about statistics or computational linguistics.

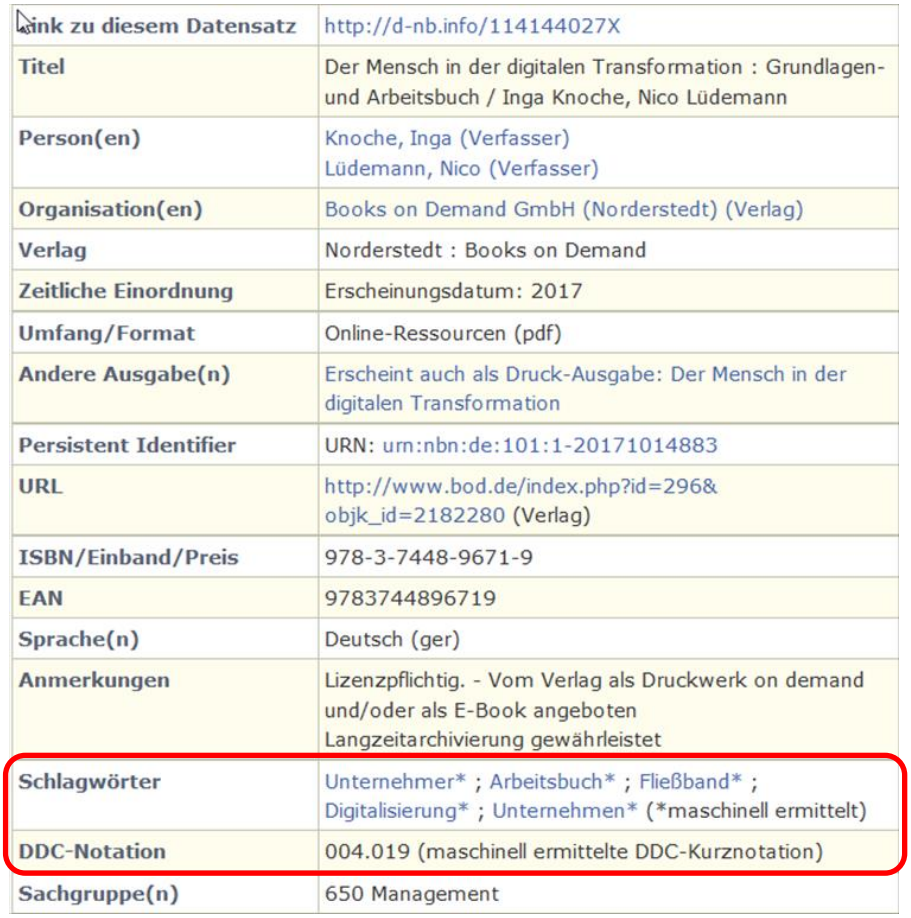
It is evident that intellectual subject cataloguing will always be required, be it for material for which acceptable results cannot be reached with automated procedures, be it for providing training material and material for evaluation of automated procedures, be it for the curation of the authority file. The smart indentation of both intellectual and automated methods is one of the major tasks in the years to come. This includes the development of smart tools and interfaces that also support intellectual subject cataloguing in an efficient way.

Data provenance

Automated generation of (subject) metadata goes along with the requirement of marking these data on the field level. The mode and process of creation of data, time-stamps, etc. is

information relevant for data analysis, the steering of workflows and quality management, but also for the use and re-use of the data in catalogues. When consulting the online catalogue, users should be able to see whether data were generated automatically. DNB's data customers should be able to decide whether and how they integrate automatically generated date into their databases. The transfer of provenance information is done via MARC21 field 883.

It is expected that provenance information on metadata will become more important in the future as library metadata become more dynamic. A catalogue record may consist of data elements of heterogeneous origin, which may updated or altered multiple times. Provenance information allows traceability.



Link zu diesem Datensatz	http://d-nb.info/114144027X
Titel	Der Mensch in der digitalen Transformation : Grundlagen- und Arbeitsbuch / Inga Knoche, Nico Lüdemann
Person(en)	Knoche, Inga (Verfasser) Lüdemann, Nico (Verfasser)
Organisation(en)	Books on Demand GmbH (Norderstedt) (Verlag)
Verlag	Norderstedt : Books on Demand
Zeitliche Einordnung	Erscheinungsdatum: 2017
Umfang/Format	Online-Ressourcen (pdf)
Andere Ausgabe(n)	Erscheint auch als Druck-Ausgabe: Der Mensch in der digitalen Transformation
Persistent Identifier	URN: urn:nbn:de:101:1-20171014883
URL	http://www.bod.de/index.php?id=296&objk_id=2182280 (Verlag)
ISBN/Einband/Preis	978-3-7448-9671-9
EAN	9783744896719
Sprache(n)	Deutsch (ger)
Anmerkungen	Lizenzpflichtig. - Vom Verlag als Druckwerk on demand und/oder als E-Book angeboten Langzeitarchivierung gewährleistet
Schlagwörter	Unternehmer* ; Arbeitsbuch* ; Fließband* ; Digitalisierung* ; Unternehmen* (*maschinell ermittelt)
DDC-Notation	004.019 (maschinell ermittelte DDC-Kurznotation)
Sachgruppe(n)	650 Management

Fig. 6: Labeling of automatically generated subject headings and DDC notations in DNB's OPAC

Conclusion

It remains one of the strategic priorities of the DNB to make use of the possibilities of automated cataloguing. The experience of the past years regarding the field of automated indexing and classification has shown that it requires substantial effort to develop procedures which work for a collection that is as diverse in types and topics of publications as DNB's collection. Nevertheless, the success achieved is reason to continuously work on the improvement of automated procedures. One of the advantages of automated cataloguing is the possibility to re-process large amounts of publications in case better methods are available. The goal is to provide as good a subject access as possible to as many publications as possible.

References

Junger, Ulrike: *Can Indexing Be Automated? The Example of the Deutsche Nationalbibliothek.* In: Cataloging & Classification Quarterly, 52 (2014), p. 102-109. Online: http://dx.doi.org/10.1080/01639374.2013.854127.

Gömpel, Renate; Junger, Ulrike; Niggemann, Elisabeth: *Veränderungen im Erschließungskonzept der Deutschen Nationalbibliothek.* In: Dialog mit Bibliotheken, 22 (2010) 1, p. 20 - 22. Online: http://nbn-resolving.de/urn:nbn:de:101-2011012858.

Junger, Ulrike; Schwens, Ute: *Die inhaltliche Erschließung des schriftlichen kulturellen Erbes auf dem Weg in die Zukunft.* In: Dialog mit Bibliotheken, 29 (2017) 2, p. 4 – 7. Online: http://nbn-resolving.de/urn:nbn:de:101-20170929367.

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

Mödden, Elisabeth: *Inhaltserschließung im Zeitalter von Suchmaschinen und Volltextsuche.* In: b.i.t online 21 (2018) 1, p. 47-51. Online: https://www.b-i-t-online.de/heft/2018-01-interview-moedden.pdf.

Mödden, Elisabeth; Schöning-Walter; Christa; Uhlmann, Sandro: *Maschinelle Inhaltserschließung in der Deutschen Nationalbibliothek.* In: BuB 1 (2018), p. 30-35.

Uhlmann, Sandro: *Automatische Beschlagwortung von deutschsprachigen Netzpublikationen mit dem Vokabular der Gemeinsamen Normdatei (GND).* In: Dialog mit Bibliotheken 25 (2013) 2, p. 26-36. Online: http://nbn-resolving.de/urn:nbn:de:101-20140305238.

Biographical note - Ulrike Junger, Head Domain Acquisition and Cataloguing, Deutsche Nationalbibliothek, Frankfurt/Main, Germany.

1995-2001: Subject librarian and editor for the German subject heading authority file, University and State Library Göttingen, Germany
2001-2009: Head of Academic Services, then Head German Union Catalogue of Serials, Berlin State Library, Berlin, Germany
Since 2009: Head Department of Subject Cataloguing, now Head Domain Acquisition and Cataloguing, German National Library, Frankfurt/Main, Germany