Automatic Analysis of Dewey Decimal Classification Notations

Ulrike Reiner

Verbundzentrale des Gemeinsamen Bibliotheksverbundes (VZG)
37077 Göttingen, Germany
ulrike.reiner@gbv.de

Abstract. The Dewey Decimal Classification (DDC) was conceived by Melvil Dewey in 1873 and published in 1876. Nowadays, DDC serves as a library classification system in about 138 countries worldwide. Recently, the German translation of the DDC was launched, and since then the interest in DDC has rapidly increased in German-speaking countries.

The complex DDC system (Ed. 22) allows to synthesize (to build) a huge amount of DDC notations (numbers) with the aid of instructions. Since the meaning of built DDC numbers is not obvious - especially to non-DDC experts - a computer program has been written that automatically analyzes DDC numbers. Based on Songqiao Liu's dissertation (Liu (1997)), our program decomposes DDC notations from the main class 700 (as one of the ten main classes). In addition, our program analyzes notations from all ten classes and determines the meaning of every semantic atom contained in a built DDC notation. The extracted DDC atoms can be used for information retrieval, automatic classification, or other purposes.

1 Introduction

While searching for books, journals, or web resources, you will often come across numbers such as "025.1740973", "016.02092", or "720.7073". What do they mean? Librarian professionals will identify these strings as numbers (notations) of the Dewey Decimal Classification (DDC), which is named after its creator, Melvil Dewey. Originally, Dewey designed the classification for libraries, but in the meantime DDC has also been discovered for classifying the web or other resources. The DDC is used, among others, because it has a long-standing tradition and is still up to date: in order to cope with scientific progress, it is currently under development by a ten-member international board (the Editorial Policy Committee, EPC). While the first edition, which was published in 1876, only comprised a few pages, the current 22nd edition of the DDC spans a four-volume work with almost 4,000 pages. Today, the DDC contains approx. 48,000 DDC notations and about 8,000 instructions. The DDC notations are enumerated in the schedules and tables of the DDC.

With the aid of the instructions mentioned above, human classifiers can build new (so-called) synthesized notations (numbers) if these are not specifically listed in the DDC schedules. This way, an enormous amount of synthesized DDC notations has been built intellectually over the last 130 years. These mostly unused notations are contained in library catalogues - like a hidden treasure. They can be considered as belonging to the "Deep Lib", one of the subsets of the "Deep Web" (Bergman (2001)). Can these notations be made accessible for information retrieval purposes with reasonable effort?

Our answer to this question consists in the automatic analysis of notations of the DDC. The analysis program written in the pattern scanning and processing language "gawk" (http://www.gnu.org/software/gawk/) determines all DDC notations (together with their corresponding captions) contained in a synthesized (built) DDC notation. Before we go into details of the automatic analysis of DDC notations in section 3, section 2 provides the basis for the analysis. In section 4, the results as well as possible applications are presented.

2 DDC Notations

Notations play an important role in the DDC:

"Notation is the system of symbols used to represent the classes in a classification system. ... The notation provides a universal language to identify the class and related classes, regardless of the fact that different words or languages may be used to describe the class." (http://www.oclc.org/dewey/versions/ddc22/intro.pdf)

The following picture serves as an example for the aforesaid. Class C is represented by the notation 025.43 or, respectively, by the captions of three different languages:

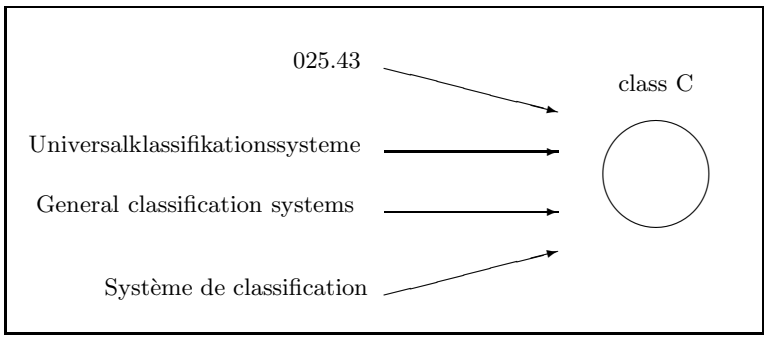

Fig. 1. Class C represented by notation 025.43 or by several captions

The DDC notations interrelate with hierarchy and structure in the following way:

"Hierarchy in the DDC is expressed through structure and notation. ... Structural hierarchy means that all topics (aside from the ten main classes) are part of all the broader topics above them. The corollary is also true: whatever is true of the whole is true of the parts. This important concept is called hierarchical force. ... Notational hierarchy is expressed by length of notation. Numbers at any given level are usually subordinate to a class whose notation is one digit shorter; coordinate with a class whose notation has the same number of significant digits; and superordinate to a class with numbers one or more digits longer." (http://www.oclc.org/dewey/versions/ddc22/intro.pdf)

In compliance with the DDC system, the automatic analysis of notations (numbers) of the DDC is carried out in the VZG (*V*erbund*Z*entrale des *G*emeinsamen Bibliotheksverbundes) project Colibri (*CO*ntext generation and *LI*nguistic tools for *B*ibliographic *R*etrieval *I*nterfaces). The goal of this project is to enrich title records on the basis of the DDC to improve retrieval. The analysis of DDC notations is conducted under the following research questions (which are also posed in a similar way in Liu (1993), p. 18):

Q1: Is it possible to automatically decompose molecular DDC notations into atomic DDC notations?

Q2: Is it possible to improve automatic classification and retrieval by means of atomic DDC notations?

We define the terms "atomic DDC notation" and "molecular DDC notation" (while a DDC notation is considered as a string, i.e., an ordered sequence of symbols) as follows:

Atomic DDC notation:
An atomic DDC notation is a semantically indecomposable string that represents a DDC class.

Molecular DDC notation
A molecular DDC notation is a string that is syntactically decomposable into dno_atoms.

General remarks: (1) We use the term "molecular DDC notation" instead of "synthesized DDC notation" to emphasize the decomposition into dno_atoms (cf. section 3). (2) We abbreviate "DDC notation" as "dno", "atomic DDC notation" as "dno_atom", "molecular DDC notation" as "dno_mol", "caption" as "cap", "schedule notation" as "schedno", and "table notation" as "tabno". (3) Technical terms (dno_atom, dno_mol, dno, cap, etc.) with appended "s" are to be understood as the respective terms' plural forms.

DDC notations can be found at several places in the DDC. In DDC summaries, the notations for the main classes (or tens), the divisions (or hundreds), and the sections (or thousands) are enumerated. Other notations are listed in the schedules ("DDC schedule notations") or (internal) tables ("DDC table notations"). DDC schedules is "the series of DDC numbers 000-999, their headings (captions), and notes." (Mitchell (1996), p. lxv). A DDC table is "a table of numbers that may be added to other numbers to make a class number appropriately specific to the work being classified" (Mitchell (1996), p. lxv). Further notations are contained in the "Relative Index" of the DDC. The frequency distributions of schedule (table) notations are shown in Fig. 2 (Fig. 3), while schedno0 is short hand for DDC schedule notations beginning with 0, schedno1 for DDC schedule notations beginning with 1, etc. The captions for the main classes are: 000: Computer science, information & general works; 100: Philosophy & psychology; 200: Religion; 300: Social sciences; 400: Language; 500: Science; 600: Technology; 700: Arts & recreation; 800: Literature; 900: History & geography. As illustrated by Fig. 2, DDC notations are not distributed uniformly: the most schednos can be found in the class "Technology", followed by the notations in the class "Social sciences". The fewest notations belong to the class "Philosophy & psychology". With regard to the tabnos (Fig. 3), the 7,816 Table 2 notations ("Geographic Areas, Historical Periods, Persons") stand out, whereas, in contrast, the quantities of all other tabnos are comparatively small (Table 1: Standard Subdivisions; Table 3: Subdivisions for the Arts, for Individual Literatures, for Specific Literary Forms; Table 4: Subdivisions of Individual Languages and Language Families; Table 5: Ethnic and National Groups; Table 6: Languages).

As mentioned before, DDC notations, which are not explicitly listed in the schedules, can be built by using DDC instructions. This process is called "notational synthesis" or "number building". Its results are synthesized DDC notations (dno_mols) that usually only DDC experts are able to interpret. But with the aid of our program component vc_day (*v*zg *c*olibri_*d*dc number *a*nal*yz*er), the meaning of dno_mols is revealed and the determined dno_atoms can be used, among others, to answer question Q2. The state of the art of the automatic analysis of DDC notations and the program components vc_day, vc_KB (*v*zg *c*olibri_*K*nowledge *B*ase), the dno input files for vc_day (in_gvk_all, in_liu_t), and the vc_day output files (vc_daygram: DDC analysis diagram and vc_dayset: DDC analysis set of dnos or captions) are subject of the next sections. When we speak of program components, we want to make clear that they belong to the main program vc_ds (*v*zg *c*olibri_*s*earch *s*ystem, cf. Fig. 4), which will not be discussed here (the dotted lines) but can be found elsewhere (Reiner (2005), Reiner (2007a), and Reiner (2007b)).

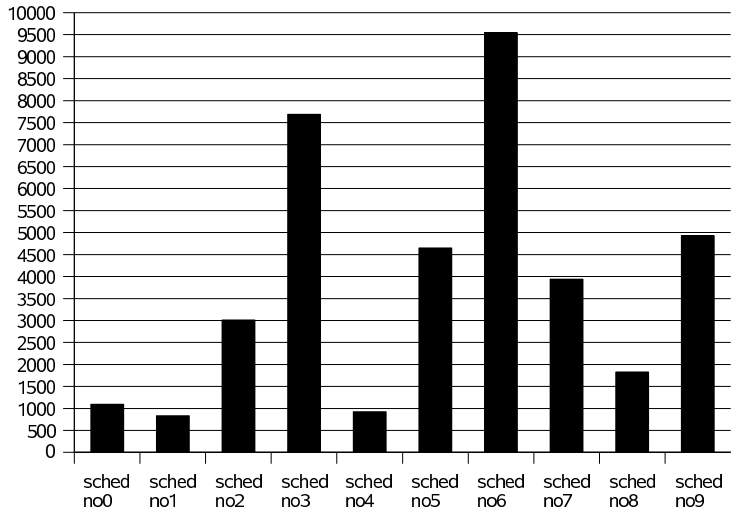

Fig. 2. Frequency distribution of DDC schedule notations

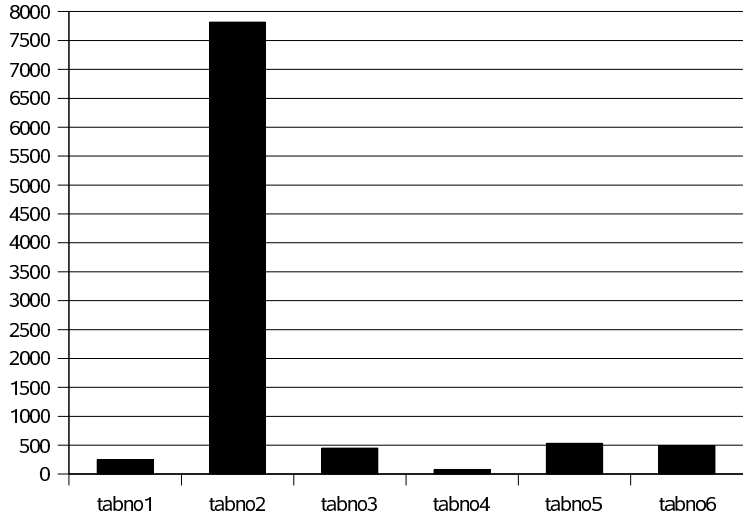

Fig. 3. Frequency distribution of DDC table notations

3 Automatic Analysis of DDC Notations

in_gvk_all. The GBV Union Catalog *GVK* (*G*emeinsamer *V*erbund*K*atalog, http://gso.gbv.de/) contains 3,073,423 intellectually DDC-classified title records (status: July, 2004). A few records have more than one DDC notation assigned

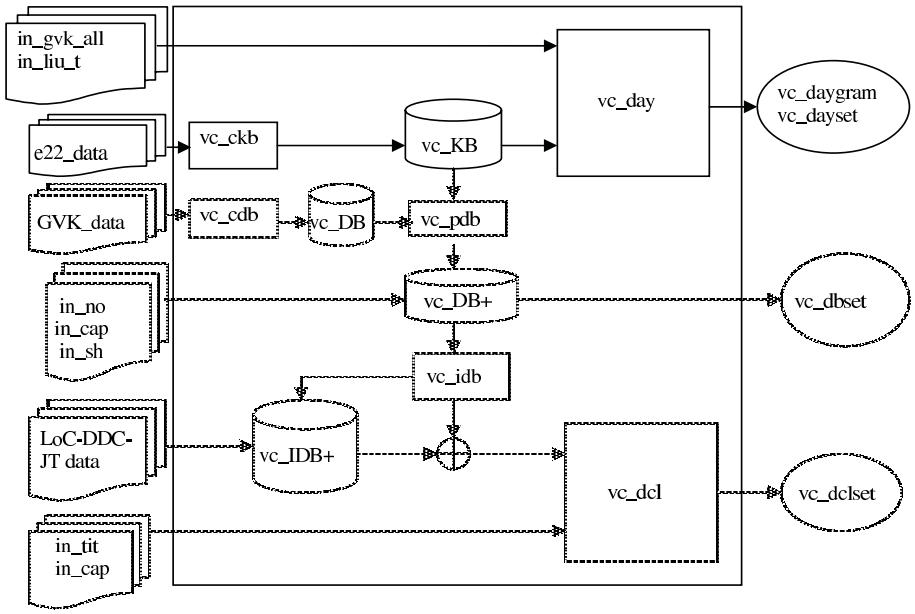

Fig. 4. System architecture of the DDC search system

to them, hence 3,339,717 DDC notations are available for the automatic analysis. Some of the strings that are stored in the field for DDC notations are incorrect DDC notations such as "#363.7", "980..711", or "081 s" as they, e.g., contain (special) characters or wrongly placed Dewey dots. Other strings contain admissible characters as segmentation marks (slash mark "/" or prime mark " ' " or other indicators). After the automatic elimination of segmentation marks, obviously incorrect DDC notations (3.8 per cent of all DDC notations), and duplicate DDC notations, a total of 466,134 different dnos is available for the automatic analysis of DDC notations. This set of all GVK dnos ("in_gvk_all") serves as input data for the analysis program. Most (189,246) dnos of "in_gvk_all" belong to dno3* (dno3* is short hand for DDC notations beginning with 3), followed by dno9* (62,115), dno7* (52,632), dno6* (51,704), dno5* (33,649), dno0* (23,946), dno2* (20,888), dno8* (20,678), dno4* (6,680), and dno1* (4,596). The arity of dnos of "in_gvk_all" is Gaussian distributed with a maximum at 10, i.e. most dnos have approx. arity 10, the shortest dno has arity 1, the longest dno has arity 29.

in_liu_t. At the beginning of the work we didn't analyze the dnos of the file "in_gvk_all" but those of the test file "in_liu_t". The file was created by numbering consecutively all 600 dnos that are given in the Appendixes C.-E. of Liu (1993) and inputting them into the file "in_liu_t" (in: input; t: test). The 600 dnos within the subject scope "Arts & recreation" were randomly selected from the OCLC database by Liu. He made a restriction on dno7* be-

cause of limited time and resources in his dissertation study. "in_liu_t" (analog in_gvk_all) has the form:

in_liu_1
700.23
in_liu_2
700.90440747471
...
in_liu_600
799.26092

After 14 years (Liu's dissertation was published in 1993), 36 DDC notations are out of date because of relocations and discontinuations.

vc_KB. As a member of the Consortium DDC German, we have access to the machine-readable data of the 22nd edition of the DDC system (e22_data in Fig. 4). e22_data incorporates the expert knowledge of the DDC system. The English electronic web version is available as WebDewey (http://connexion.oclc.org/), the German pendant as MelvilClass (http://services.ddc-deutsch.de/melvilclass-login). The program component vc_ckb (*v*zg *c*olibri_*c*reate *k*nowledge *b*ase) creates vc_KB (*v*zg *c*olibri_*K*nowledge *B*ase). Only those data of e22_data are extracted that are necessary for the program vc_day. Due to efficiency reasons, natural language phrases are ignored, same as xml lines (e22_data is an xml file) that are unessential for analysis purposes, e.g., xml tags "<lnk idr>" ("identifier for record"), "<mtz>" ("column in manual text table"), or "<dtu>" ("date updated"). Hence, the data structure of e22_data is transformed into a convenient data structure for analysis. In vc_KB, the data dno, descriptor, and descriptor value(s) are stored in consecutive fields, while facts and rules are represented in a very similar way and '#' serves as field separator (there can be empty fields as in the 025.17 line):

T1–093-T1–099+021#<ba4r2>#Statistics
025.17#<na1r1>##025.17#025.341-025.349#025.34#####
025.344#<hat>#Electronic resources

The three example vc_KB lines should be read as follows: *Fact*: T1–093-T1–099+021 has the caption "Statistics". *Rule*: Add to base number 025.17 the numbers following 025.34 in 025.341-025.349. *Fact*: 025.344 has the caption "Electronic resources". Again the e22_data xml tags are given in angle brackets: "<ba4>" ("beginning of add table, (all of table number)"), "<na1>" ("add note (part of schedule number)"), and "<hat>" ("hierarchy at class"). "r1" and "r2", which follow the e22_data xml tags "<na1>" and "<ba4>", respectively, stand for the first two macro rules (see below).

vc_KB contains 48,067 facts and 8,033 rules (2 MB file size). [*Note*: For ef-

ficiency reasons again, we also built vc_KB_spli (spli: splitting the span of numbers), which contains 121,035 facts and 80,324 rules (15 MB file size)]. The 8,033 (80,324) rules can be generalized to macro rules. While Liu (1993) defined 17 (macro) rules for decomposition for class 700, we defined 25 macro rules for all DDC classes. This amount could even be reduced to one regular expression (which, however, is not done in our program due to usability reasons): *egrep -i "add to base number.*(notation | the numbers following.*[notation]*) | add to each subdivision identified by.*(as follows | notation | the numbers following | as instructed (at | under)) | (add [to]* | add to base number).*as instructed (at | under) | (add.*notation.*from.*table.*[under]*) | add [to]*.*(the [historical period | period division]*numbers following | notation)" e22_data.*

vc_day. After initializing variables, the analyzer vc_days (vc_days is the "splitting" version of vc_day for vc_kb_spli) reads the knowledge base vc_kb_spli, and, triggered by one dno or more dnos, analyzing of dnos is performed. The number of correct and incorrect dnos is counted. For a dno, there are two phases to the analyzing process including: phase 1: determining the facts from left to right, phase 2: determining the facts via rules from left to right. After checking which output format has to be printed, the result is printed as a so-called vc_daygram (DDC analysis diagram) or as a vc_dayset (DDC analysis set of dnos or captions). After all dnos have been analyzed, the number of totally/partially analyzed dnos is printed. There are different reasons for a partially analyzed dno: either the implementation of vc_days is incorrect/still incomplete or the dno is incorrectly synthesized or the DDC system itself is incorrect (e.g., vc_days finished with an incomplete analysis at several dno_mols because of one typographical error of a span of numbers while analyzing in_gvk_all).

4 Results and Applications

Results. To illustrate the results, we give a comparison between Liu's and our result (as vc_daygram) for the dno_mol "in_liu_37":

Liu (1993), pp. 99–100
"720.7073 has been decomposed as follows:
720: Architecture
0707: Geographical treatment
73: United States
The title of this book is:
#aVoices in architectural education: #bcultural politics and
The subject headings for this book are:
#aArchitecture #xStudy and teaching #zUnited States.
#aArchitecture and state #zUnited States."

Reiner (2007), p. 49
`720.7073` <ul-liu/liu_37_to_analyze; length: 8>
`7-------` Arts & recreation <hatzen>
`72------` Architecture <hatzen>
`720-----` Architecture <hat>
`--0.7---` Education, research, related topics <T1–07>
`--0.707-` Geographic treatment <T1–0707>
`---.--7-` North America <na4r7span:T1–0701-T1–0709:T2–7>
`---.--73` United States <na4r7span:T1–0701-T1–0709:T2–73>

The information given in angle brackets stands for: "<hatzen>" is the concatenation of "<hat>" ("hierarchy at class") and "<zen>" ("zen built entry (main tag)"), "<T1–" (table 1), "<T2–" (table 2), "<na4>" ("add note (add of table number)"), "r7" (macro rule 7), "span" (span of numbers), and ":" (delimiter). As you can see, while Liu decomposes the synthesized dno into three chunks, our vc_daygram shows the finest possible analysis of dno_mol. The fine analysis provides the advantage of uncovering additional captions: "Arts & recreation", "Architecture" as well as "North America", and also "Education, research, related topics" (notice that the term "education" is part of the title of the book). As a result of Liu's approach, the latter caption is missed due to his procedure for decomposing numbers:

"1. Match the number to be decomposed against the Schedule by dropping digits on the right. If the complete number is found in the Schedule, the number is not a synthesized number and decomposition is complete. If no match is made after all digits are dropped, set the number aside as containing a potential error. 2. When a match is made, search the entry in the Schedule for an Add Note. If one is found, determine which note type it is and apply the rule defined for that type. If none is found, apply appropriate rules for Standard Subdivisions. 3. Repeat the above two steps for the remaining digits, but in repetitions, in the first step the number may be searched against the Tables rather than the Schedule." (Liu (1993), pp. 40–41)

As it was our aim to determine every semantic atom, we decided to follow a different approach than Liu (1993). In contrast to his one-phase approach from right to left, our algorithm is a two-phase approach from left to right in both phases. The strength of our approach can also be seen in dno_mols like the following (additionally to the caption "City planners", the caption "Persons" is obtained by phase 2):

Reiner (2007)
`711.4092` <liu_26_to_analyze; length: 8>
`7-------` Arts & recreation <hatzen>
`71------` Landscaping & area planning <hatzen>

711----- Area planning (Civic art) <hat>
711.4--- Local community planning (City planning) <hat>
711.4092 City planners <hatien>
---.-09- Historical, geographic, persons treatment <T1–09>
---.-092 Persons <T1–092>

"<hatien>" is the concatenation of "<hat>" and "<ien>" ("built schedule entry (main tag)"). A vc_daygram contains analysis and synthesis information: 1. the DDC notation to be analyzed (dno_mol); 2. an identifier (name) and the length of dno_mol; 3. the sequence and position of the digits within dno_mol; 4. the Dewey dot at position 4; 5. the relevant parts of dno_mol for each analysis step (dno_atom); 6. the corresponding caption for every dno_atom; 7. the parts irrelevant for the respective analysis step marked with "-"; 8. the type of the applied facts and rules that appear in angle brackets. In case it has been explained how to read the given information mentioned in 8., every synthesis step can be reproduced (in expert systems, this capability is achieved by the explanation component).

While vc_daygrams are intended for human experts, the vc_daysets can be used for data transfer. Currently, we distinguish three kinds of vc_daysets (if needed, other formats/forms are certainly possible): vc_dayset_dno_cap, vc_dayset_fine, and vc_dayset_mab2, which are shown for dno_mol "in_liu_37":

---

vc_dayset_dno_cap: 720.7073 <liu_37_to_analyze>

7;Arts & recreation
72;Architecture
720;Architecture
T1–07;Education, research, related topics
T1–0707;Geographic treatment
T2–7;North America
T2–73;United States

---

vc_dayset_fine: 720.7073 <liu_37_to_analyze>

---

liu_37:720.7073;7;72;720;T1--07;T1--0707;T2--7;T2--73

---

vc_dayset_mab2: 720.7073 <liu_37_to_analyze>

705aˆ_a720.7073ˆ_p72ˆ_cT1–070ˆ_f0707ˆ_g73

This way, all 600 dnos of in_liu_t have been analyzed, printed, and compared

accordingly with the results of Liu (1993). It turns out that Liu's result can be reproduced. Minor differences result from typing/printing errors in his dissertation and the usage of different DDC editions. While Liu (1993) used the 20th Edition, we used the 22nd Edition of the DDC. vc_days analyzes a further 27 dnos of the 600 dnos that Liu could not decompose because of his restriction to the class 700. As far as the analysis of in_gkv_all (466.134 dnos of all DDC classes) is concerned, currently 297,782 (168,352) dnos can be totally (partially) analyzed, i.e. 63.9 per cent (36.1 per cent) are totally (partially) analyzed. In some DDC classes, the analyzing degree is even higher, which means that, e.g., 87 per cent of the 51,704 DDC notations of the class "Technology" (600) can be totally analyzed.

Applications. Analyzing results can serve different purposes, which have already been mentioned in Liu (1993), pp. 66: 1. dno_atoms could be used to improve recall and precision of information retrieval systems. 2. dnos could be used as switching language for multilingual retrieval. 3. On the basis of vc_daygrams, DDC tutorials, or expert systems could be developed to support teaching or quality control of notational DDC synthesis.

References

BERGMAN, Michael K.: The Deep Web: Surfacing Hidden Value. *The Journal of Electronic Publishing, Volume 7, Issue 1, August 2001.* Online: http://www.press.umich.edu/jep/07-01/bergman.html

MITCHELL, J. (ed.) (1996): *Dewey Decimal Classification and Relative Index. Ed. 21, Volumes 1-4.* Forest Press, OCLC, Inc., Albany, New York, 1996. (http://connexion.oclc.org/).

LIU, Songqiao (1993): *The Automatic Decomposition of DDC Synthesized Numbers.* Ph.D. diss., University of California, Graduate School of Library and Information Science, Los Angeles, 1993.

REINER, Ulrike (2005): *VZG-Projekt Colibri - DDC-Notationsanalyse und -synthese.* September 2004 - Februar 2005. VZG-Colibri-Bericht 2/2004. Verbundzentrale des Gemeinsamen Bibliotheksverbundes (VZG), Göttingen, 2005.

REINER, Ulrike (2007a): *Automatische Analyse von Notationen der Dewey-Dezimalklassifikation.* 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications - Librarian Workshop: Subject Indexing and Library Science. March 7-9, 2007, Freiburg i. Br., Germany. (http://www.gbv.de/vgm/info/biblio/01VZG/06Publikationen/2007/pdf/pdf_2835.pdf).

REINER, Ulrike (2007b): *Automatische Analyse von DDC-Notationen und DDC-Klassifizierung von GVK-Plus-Titeldatensätzen.* Workshop zur Dewey-Dezimalklassikation "DDC-Einsichten und Aussichten 2007". March 1, 2007, SUB Göttingen, Germany. (http://www.gbv.de/vgm/info/biblio/01VZG/06Publikationen/2007/pdf/pdf_2836.pdf).