VZG-Projekt Colibri

DDC-Notationsanalyse und -synthese
September 2004 - Februar 2005

Ulrike Reiner

VZG-Colibri-Bericht 2/2004

Verbundzentrale des
Gemeinsamen Bibliotheksverbundes (VZG)
Platz der Göttinger Sieben 1
D-37073 Göttingen
http://www.gbv.de/

Name des Dokuments

colibri03.doc (Colibri-Aufnahme aus: http://www.interlog.com/barrow/coperum.jpg)

Inhalt

1. Einleitung 3
2. DDC-Notation 4
3. DDC-Notationsanalyse und –synthese 5
3.1 Einführung 5
3.2 Automatische Zerlegung von DDC-Notationen nach Liu (DND) 8
3.3 VZG-Colibri-Ansatz zur Notationsanalyse 15
3.3.1 Problemstellung 15
3.3.2 DDC-Suchsystem (vc_ds) mit Analysekomponente (vc_day) 15
3.3.3 Algorithmus zur DDC-Notationsanalyse 27
4. Vergleich der Ergebnisse der Notationsanalysen (DND, vc_day) 37
5. Zusammenfassung und Ausblick 42

Versionen

colibri03-04-09-07.doc	Entwurf -
colibri03-05-02-14.doc	Fortführung
colibri03-05-04-11.doc	Endfassung (terminologische und kleinere Korrekturen)

Dank

Ich danke Prof. Konrad (TU Berlin, Fakultät Elektrotechnik und Informatik, Fachgebiet Wissensbasierte Systeme) für sein Interesse an der hier behandelten Thematik und für seine kritische Durchsicht dieses Berichtes. Seine nützlichen Anregungen haben in diesen Bericht Eingang gefunden. Mein Dank geht auch an Frau Dr. Heidrun Alex (Die Deutsche Bibliothek, Projekt DDC Deutsch, Frankfurt am Main), die mir Hinweise zur Berichtsversion „colibri03-05-02-14.doc“ gegeben und mir die deutsche Übersetzung „Einleitung in die Dewey-Dezimalklassifikation“ („Introduction to Dewey Decimal Classification“) und „Glossar“ („Glossary – DDC Edition 22“) zur Verfügung gestellt hat. Ihre Hinweise und der in den deutschen Texten verwendete Sprachgebrauch wurden in der Endfassung mitberücksichtigt.

1. Einleitung

1876 wird das "Dewey System" anonym unter dem Titel „A Classification and Subject Index for Cataloguing and Arranging the Books and Pamphlets of a Library" in Amherst, Massachusetts, publiziert[1]. Die nach Melville Louis Kossuth Dewey - kurz Melvil Dewey - benannte Dewey-Dezimalklassifikation (DDC) revolutioniert 1920 die Bibliothekswelt[2]. Seine Überlegungen führen zu einem Klassifikationssystem, das die logische und physische Datenunabhängigkeit bei Datenbanksystemen antizipiert. Das ortsunabhängige, inhaltsbezogene DDC-System überwindet die bis dahin verwendeten, ortsgebundenen Klassifikationssysteme[3]. Sein Leitgedanke ist, das gesamte menschliche Wissen auf eine hierarchische Klassenstruktur abzubilden. Das DDC-System basiert auf drei Grundprinzipien:

1. Reihenfolge der Klassifizierung zuerst nach Fachgebieten (disciplines), dann nach Themen (subjects), Gegenständen (topics) und schliesslich nach Aspekten (aspects).
2. Klassifizierung vom Allgemeinen zum Spezifischen mit Eigenschaftsvererbung (Hierarchische Vererbung, hierarchical force).
3. Widerspiegelung der strukturellen DDC-Hierarchie durch die notationale Hierarchie (notational hierarchy). Je länger eine DDC-Notation, desto spezieller die dadurch benannte Klasse.

Seit über 120 Jahren wird die Dewey-Dezimalklassifikation laufend von DDC-Experten weiterentwickelt und erscheint ca. alle 7 Jahre in zwei gedruckten Versionen: als ungekürzte Ausgabe und als Kurzausgabe (full / abridged version). Seit 1993 ist die Dewey-Dezimalklassifikation auch elektronisch verfügbar, zunächst als MS-DOS-Version, seit 2000 plattformunabhängig als „WebDewey"[4]. WebDewey enthält gegenüber der gedruckten Version (4-bändiges Werk) u. a. eine größere Mengen an Themen- und Fachgebietsbegriffen (relative index terms)[5] und mit Anweisungen erstellte (synthesized, built) DDC-Notationen (synthetische Notationen); intellektuell und maschinell (intellectual / statistical association) erstellte Zuordnungen der DDC-Fachgebietsbegriffe zu den Schlagwörtern der Library of Congress (LCSH: LC Subject Headings)[6], Zugriff auf die assoziierten LCSH-Normdatensätze und vierteljährliche Datenbankaktualisierungen. Die DDC-Kurzausgabe ist ein Klassen-Auszug aus allen DDC-Klassen der ungekürzten Ausgabe, entstanden durch rechtsseitige Längenabschneidung der DDC-Notationen an (von zentralen Katalogdiensten wie der „Decimal Classification Division of the Library of Congress" oder der „National Library of Canada" eingebrachten) semantisch[7] sinnvollen Stellen.

[1] [ChanMitchell2003], p. 4. Lois Mai Chan; Joan S. Mitchell: "Dewey Decimal Classification: Principles and Application". 3rd ed. OCLC Online Computer Library Center, Inc., Dublin, Ohio, 2003
[2] [ChanMitchell2003], p. 2
[3] [ChanMitchell2003], p. 3
[4] Die an WebDewey orientierte deutsche Variante „MelvilSearch" und „MelvilClass" sind seit April 2004 im Entstehen begriffen, s. http://services.ddc-deutsch.de/
[5] [ChanMitchell2003], p. 194. „Relative Index {Anm. d. Autorin: DDC-Register}: The index to the DDC. It is called 'Relative' because it shows the connection between subjects and the disciplines in which they appear."
[6] [http://connexion.oclc.org/] - WebDewey Explanations: "EM: Editorially mapped LCSH (intellectual); PPT: People, Places & Things (intellectual / statistical); SM: OCLC WorldCat (statistical)"
[7] im Sinne der logischen Semantik

Das DDC-Notationssystem - als ein Teil des gesamten DDC-Systems – steht im Mittelpunkt dieser Arbeit, Kenntnis über das DDC-System wird in dieser Arbeit vorausgesetzt. Die nachfolgenden Kapitel behandeln unterschiedliche Aspekte der DDC-Notation wie Bildung, Zerlegung und erste Überlegungen zum Aufbau einer formalen DDC-Sprache. Grundlage bildet die 22. WebDewey-Ausgabe - sofern nicht anders erwähnt.

2. DDC-Notation

Unter Notation(ssystem) wird hier ein Zeichensystem verstanden, das die Klassen in einem Klassifikationssystem repräsentiert. In der Dewey-Dezimalklassifikation ist das Notationssystem (DDC numbering system) durch arabische Ziffern[8] von „0“ bis „9“ (DDC numbers) realisiert. Dadurch wird eine universelle Sprache bereitgestellt, um Klassen ausdrucks- und sprachunabhängig eindeutig zu benennen[9]. Durch eine Notation wird einer Klasse eine eindeutige Bedeutung zugeordnet und die Beziehung zu anderen Klassen festgelegt. Somit repräsentiert die Notation die Klassenstruktur, sie wird durch die Notation linearisiert. Alle korrekten (zulässigen) DDC-Notationen sind entweder in der DDC-Haupttafel (schedules) oder in den sechs DDC-Hilfstafeln (tables) aufgezählt. Bspw. steht die Klassennotation (class number) „635.7” (Aromatic and sweet herbs) in der DDC-Haupttafel und die Tafelnotation (table number) „T2--143“ (Elevations) in Hilfstafel 2. Weitere Beispiele (dno1, ..., dno8) sind:

```
dno1: 500         Natural sciences and mathematics
dno2: 549.114     Minerals in rocks
dno3: 549.7       Other minerals
dno4: 549.782     Calcite group
dno5: 581.7538    Alpine plants
dno6: 600         Technology (Applied sciences)
dno7: 635.9528    Alpine plants
dno8: 688.76522   Mountain climbing--equipment technology
```

Abb. 2.1 Beispiele für DDC-Klassen (8 DDC-Notationen mit DDC-Klassenbenennungen)

In WebDewey (22. Ausgabe, Juli 2003) werden folgende DDC-Notationen[10] unterschieden:

	Beispiel
Klassennotation (Class number)	004.67
Synthetische Klassennotation (Built class number)	B598.176
Zentrierter Eintrag, Bereich von Klassennotationen, in Seitenmitte stehend (Centered entry)	C930-990
Optionale Klassennotation, in runder Klammer stehend, für lokale Belange (Optional number)	(921)
Stillgelegte Notation, in eckiger Klammer stehend, nicht zu verwenden (Bracketed number)	[104]
Hilfstafelnotation (Table number)	T1--071
Synthetische Hilfstafelnotation (Built table number)	T4--092
Zentrierter Eintrag, Bereich von Hilfstafelnotationen, in Seitenmitte stehend,(Table centered entry)	T2--4-T2--9
Klassennotation, in der Praxishilfe stehend (Manual number)	M361-365

Abb. 2.2 Arten von DDC-Notationen

[8] ggf. durch weitere Zeichen und zusätzliche Hilfszeichen, wie „ ’ “ bzw. „ / “ oder „A“, „B“, ...

[9] [http://connexion.oclc.org/] Introduction to the Dewey Decimal Classification – Edition 22. OCLC Connexion - Help Center: „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.”

[10] [http://connexion.oclc.org/] Help Center - WebDewey record – DDC number

Synthetische DDC-Notationen sind in unterschiedlichen Bibliographien verzeichnet und werden an mehreren Institutionen erzeugt[11], z. B.: Australian National Bibliography, British National Bibliography, Indian National Bibliography, British Library, Library of Congress, OCLC (Online Computer Library Center), RLG (The Research Libraries Group, Inc.) und CIP (Cataloguing In Publication).

Einer vorgelegten DDC-Notation wie z. B. der DDC-Notation "688.76522" (Mountain climbing--equipment technology, s. „dno8" aus Abb. 2.1) kann - ohne weiteren Hinweis oder Expertenwissen - nicht angesehen werden, ob es sich um (k)eine synthetische DDC-Notation handelt, hierzu ist eine DDC-Notationsanalyse mithilfe des DDC-Systems notwendig. Diese wird üblicherweise intellektuell vorgenommen: ein Blick in den WebDewey-Datensatz zur DDC-Notation "688.76522" läßt erkennen, daß es sich um eine synthetische DDC-Notation (Built class number) handelt. Wie die DDC-Notation "688.76522" konstruiert wurde, ist dort jedoch nicht direkt zu entnehmen, sondern es muss (im einfachsten Fall) in der Klassenhierarchie aufgestiegen und den Synthese-Anweisungen (add notes) bzw. Synthesehinweisen (number-built notes) zur DDC-Notationssynthese nachgegangen werden.

Möglichkeiten und Nutzen zur Automatisierung einer DDC-Notationsanalyse bzw. DDC–Notationssynthese werden im Weiteren untersucht.

3. DDC-Notationsanalyse und –synthese

3.1 Einführung

Das DDC-System enthält Elemente einer aufzählenden Klassifikation (Menge von DDC-Notationen) als auch Elemente einer Facettenklassifikation (Menge von DDC-Notationsteilen). Gegenüber den Vorgängerversionen enthält die 22. Auflage wesentlich mehr DDC-Notationen und Möglichkeiten zur Notationsbildung, wodurch sich die Komplexität des DDC-Systems im Laufe der Jahre erhöht hat.

Unter DDC-Notationssynthese (notational synthesis, number building) wird der Prozess verstanden, DDC-Notationen zu verlängern, z. B. an eine DDC-Grundnotation (base number)[12] eine (Teil-)Notation aus den DDC-Haupt-, Hilfs-, Zeit- und internen Anhängetafeln (schedules, auxiliary tables, period tables, add tables) anzufügen. Die DDC-Notationsanalyse ist – oberflächlich betrachtet[13] - der hierzu inverse Prozess. Ziel der DDC-Notationssynthese ist die Konstruktion neuer im DDC-System noch nicht enthaltener DDC-Notationen, Ziel der DDC–Notationsanalyse die Zerlegung synthetischer Notationen in ihre einzelnen Notationsbestandteile.

[11] [ChanMitchell2003], p. 52

[12] [ChanMitchell2003], p. 189. "Base number: A number of any length to which other numbers are appended." p. 74. „... the base number, which can be as brief as one digit or as long as six or seven." Eine im Rahmen dieser Arbeit vorgenommene WebDewey-Datenanalyse ergibt, dass auch längere als 7-ziffrige DDC-Grundnotationen vorkommen, z. B. "155.67182" (individual psychology)

[13] Genauer in Kapitel 3.3

Neue DDC-Notationen können auf zwei Arten gebildet werden[14]:

Anfügen von DDC-Notation(steil)en
A1.) an eine DDC-Notation aus Hilfstafel 1 (Table 1: Standard Subdivisions) , falls nicht ausdrücklich ausgeschlossen
A2.) an eine DDC-Grundnotation aufgrund von Synthese-Anweisungen (instructions)

Abb. 3.1.1 Möglichkeiten zur Bildung neuer DDC-Notationen

Beispiel 3.1.1

Nach A1.) kann aus „796.522" („Mountains, hills, rocks") z. B. die neue DDC-Notation „796.522092" („Mountaineers") gebildet werden.

Beispiel 3.1.2

Nach A2.) können an die DDC-Grundnotation „688.76"

```
688.76  Equipment for outdoor sports and games
```

Abb. 3.1.2 Beispiel einer DDC-Grundnotation

mit der DDC-Anweisung „Add-to-688.76"

```
„Add to base number 688.76 the numbers following 796 in 796.1-796.9, e.g.,
tennis rackets 688.76342; however, For skates, skateboards, skis, see
685.36; For athletic gloves and mitts, see 685.43; For camping equipment,
see 685.53. For equipment for equestrian sports and animal racing, see
688.78; For equipment for fishing, hunting, shooting, see 688.79."
```

Abb. 3.1.3 Beispiel einer DDC-Anweisung zur DDC-Notationssynthese

DDC-Notationen angefügt und damit z. B. folgende neue DDC-Notationen bzw. spezifischere DDC-Klassen gebildet werden:

```
  688.76     Equipment for outdoor sports and games (Grundnotation)
             und
  796.522    Mountains, hills, rocks (Notation aus Haupttafel)
             ergibt
688.76522    mountain climbing--equipment technology
             (konstruierte Notation)

             analog können folgende neue Klassen gebildet werden:
688.7622     Skateboarding--equipment technology
688.76334    Soccer (association football)--equipment technology
688.76342    Tennis--equipment technology
688.7651092  Hikers--equipment technology
688.76522092 Mountaineers--equipment technology
```

Abb. 3.1.4 Beispiele für neu konstruierte DDC-Notationen

[14] [ChanMitchell2003], pp. 73-80. "Chapter 7: Synthesis of Class Numbers or Practical Number Building" und [ChanMitchell2003], pp. 152-163. "Chapter 13: Number Building for Complex Subjects"

Der Prozess zur DDC-Notationssynthese umfasst mindestens folgende Schritte[15]:

S1. Auswahl aus der Menge aller in Betracht kommenden Notationskandidaten („order of preference“ und „rules“)
S2. Identifizierung der korrekten Grundnotation
S3. Lokalisierung aller hinzuzufügenden Notationen / Notationsteile („instructions“)
S4. Ermittlung der Reihenfolge der hinzuzufügenden Notationsteile („citation order“)
S5. Überprüfung, ob synthetische Notation nicht mit vorhandenen Notationen und Anweisungen konfligiert („correctness“)

Zu S1. folgt ein WebDewey-Beispiel zur Vorzugsreihenfolge (order of preference) für die DDC-Notation „641.5“ („Cooking“)[16]:

```
Preparation of food with and without use of heat.

Unless other instructions are given, observe the following table of
preference, e.g., outdoor cooking for children 641.5622 (not 641.578 ):

   Cooking for special situations, reasons, ages                  641.56
   Quantity, institutional, travel, outdoor cooking               641.57
   Money-saving and timesaving cooking                            641.55
   Cooking with specific fuels, appliances, utensils              641.58
   Cooking specific meal                                 641.52-641.54
   Cooking by specific types of persons                           641.51
   Cooking characteristic of specific geographic
     environments, ethnic cooking                                 641.59

Class menus and meal planning in 642
For cooking specific materials , see 641.6
For specific cooking processes and techniques , see 641.7
For cooking specific kinds of dishes, preparing beverages , see 641.8.
```

Abb. 3.1.5 Anmerkungen („notes“) zum WebDewey-Datensatz “641.5” (Cooking)

Weiterhin sind die in S1. genannten Regeln („rules“) zu beachten, die ggf. Vorrang vor der Vorzugsreihenfolge haben. Werden im zu klassifizierenden Werk innerhalb eines Fachgebietes mehrere Themen abgehandelt, sind die Regeln R1. - R4. zu beachten, sind mehrere Fachgebiete Gegenstand der Untersuchung, ist Regel R5. anzuwenden:

R1. Stehen in einem Werk zwei Themen in Beziehung zueinander, ist die DDC-Notation mit dem Thema zu wählen, auf das Bezug genommen wird (“rule of application”). Diese Regel hat Vorrang vor allen anderen Regeln.

R2. Werden in einem Werk zwei Themen gleichwertig behandelt und führen einander nicht gegenseitig ein, ist die kürzere der beiden DDC-Notationen zu nehmen; d.h. die DDC-Notation, die als erste in der Tafel aufgeführt ist („first-of-two-rule“). Diese Regel kann in einer Anweisung oder Anmerkung ggf. wieder aufgehoben werden.

[15] extrahiert und konkludiert aus [ChanMitchell2003] und [http://connexion.oclc.org/] “Introduction to the Dewey Decimal Classification – Edition 22. OCLC Connexion”

[16] http://connexion.oclc.org/WebZ/QUERY?sessionid=cnx03.prod.oclc.org-57120-e4b19n0d-e1z8h8&termsrch-sn%3A=641.5&format=F&next=/WebZ/html/corc/corcframe.html:entitycorcLink=/WebZ/html/corc/deweyrecordframe.html&bad=html/corc/refine-dewey.html&:entityviewingBrowse=true:entitytoprecno=1:entitycurrecno=1:numrecs=1:foo=single:dbname=DeweyDB

R3. Werden in einem Werk drei oder mehrere Themen in unterschiedlichen Unterklassen gleichwertig abgehandelt, ist die DDC-Notation der den Unterklassen direkt übergeordneten Klasse („rule of three“) auszuwählen.

R4. Bei DDC-Notationen ab der 4. Ebene („subdivisions“) sind die DDC-Notationen mit keinen oder wenigsten „Nullen“ zu bevorzugen („rule of zero“).

R5. Falls keine Regel anwendbar ist und keine DDC-Notation bevorzugt werden kann, ist folgende (Nothilfs-)Tafel als letzte Hilfe („table of last resort“) zu verwenden:

```
1. Arten von Etwas („Kinds of things“)
2. Teile von Etwas (“Parts of things”)
3. Materialien, aus denen Etwas geschaffen ist
   (“Materials from which things, kinds, or parts are made”)
4. Eigenschaften von Etwas
   (“Properties of things, kinds, parts, or materials”)
5. Vorgänge innerhalb von Etwas
   (“Processes within things, kinds, parts, or materials”)
6. Handlungen, die auf Etwas angewendet werden
   (“Operations upon things, kinds, parts, or materials”)
7. Hilfsmittel zur Darstellung solcher Handlungen
   (“Instrumentalities for performing such operations”)
```

Abb. 3.1.6 „Table of last Resort“ als Auswahlhilfe bei gleichwertigen DDC-Notationen

Die Schritte S1.- S5. enthalten semantische und syntaktische Aufgaben. Es ist zu vermuten und in der Literatur beschrieben worden[17], dass sich zumindest die Schritte S4. und S5. automatisieren lassen (syntaktische Syntheseschritte). Die Schritte S1.-S3. erfodern u.a. die Bestimmung des Inhaltes eines Werkes und dessen korrekte Repräsentation durch eine DDC-Notation (semantische Syntheseschritte), die vermutlich höchstens semi-automatisch unterstützt werden können.

Nach den einführenden Überlegungen zur Notationsanalyse und -synthese in diesem Kapitel wird die in der Literatur bislang ausführlichste Abhandlung zur DDC-Notationsanalyse im folgenden Kapitel näher beleuchtet.

3.2 Automatische Zerlegung von DDC-Notationen nach Liu (DND)

Dieses Kapitel bespricht die Arbeit von Liu1993[18]. Liu vermutet in seiner unter dem Titel „The Automatic Decomposition of DDC Synthesized Numbers" verfassten Dissertation, dass sich seine aus der DDC-Hauptklasse „700“ (Künste und Unterhaltung) gewonnenen Ergebnisse der DDC-Notationszerlegung auch auf die anderen neun DDC-Hauptklassen (mit Programmerweiterungen und -modifikationen) übertragen lassen:

[17] Quellenangaben in [Liu1993], s. folgendes Kapitel 3.2, Fussnote 24

[18] [Liu1993] Songqiao Liu: "The Automatic Decomposition of DDC Synthesized Numbers." Ph.D. diss., University of California, Los Angeles, 1993

„Since the success rate of the decomposition is more than 95% (actually 100%), it is concluded that synthesized DDC numbers can be *accurately* decomposed by computer. Although the study covers only DDC synthesized class numbers in the main class Arts (700), I strongly believe that synthesized numbers in all the DDC main classes can be accurately decomposed automatically, on the ground that the synthesizing rules for the Arts (700) are representative of the rules throughout the schedule. Synthesizing rules for the main classes Literature (800) and Language (400) may differ somewhat from those for other classes. ... thus, with some modification or augmentation, the decomposing rules that were defined by this study could be extended to all synthesized numbers in all DDC main classes. Nevertheless, it seems a logical next step to extend and apply the methodology to other DDC schedules. Future research on automatic decomposition of synthesized numbers in the other nine main DDC classes would validate and supplement the findings and conclusions of the present study.“ [Liu1993], p. 65

Abb. 3.2.1 Liu’s Schlußfolgerungen aus seiner Untersuchung zur DDC-Notationszerlegung

Da das von Songqiao Liu in seiner Dissertation erwähnte und in „MS FoxPro[19]“ von ihm entwickelte Computersystem „DND“ (Dewey Number Decomposer)[20] nicht verfügbar ist und DND-Implementierungsdetails als auch DND-Programmcode nicht dokumentiert sind[21], wird eine DDC-Notationsanalyse im Rahmen des VZG-Projektes Colibri neu entwickelt[22]. Ein Ziel innerhalb des VZG-Projektes Colibri ist es, Liu’s Zerlegungsergebnisse zur DDC-Notationsanalyse zu reproduzieren, sie mit den eigenen Ergebnissen zu vergleichen und die Analyse auf alle DDC-Klassen auszuweiten. Wenn es gelingt, jede beliebige DDC-Notation eindeutig in ihre einzelnen Bestandteile zu zerlegen, können diese für unterschiedliche Zwecke verwendet werden, wie z. B. Gewinnung von (multilingualen) Suchbegriffen für eine DDC-basierte Suche, automatische Indexierung und Klassifikation, Expertensysteme zur DDC-Notationsanalyse-/synthese, DDC-Notations-Qualitätskontrolle und -Training[23].

Während Arbeiten und Veröffentlichungen zur DDC-Notationsanalyse laut Liu rar sind, liegen einige Untersuchungen zur DDC-Notationssynthese vor[24]. Letztere benutzen zur Bildung von DDC-Notationen Information z. B. aus dem DDC-System, aus Titelaufnahmen und / oder von (un)kontrollierten Begriffen aus (natürlichsprachlichen) Texten.

Nach Liu’s Recherchen sind Wajenberg 1983[25] und Duncan & Williams 1991[26] die Ersten, die sich um eine DDC-Notationsanalyse bemühen[27]. Wajenberg’s Vorschlag beruht auf sechs intellektuell vorgenommenen zusätzlichen MARC-Unterfeldeinträgen[28], bestehend aus DDC-Notationsteilen, mit denen zerlegbare DDC-Notationen analysiert werden können.

[19] http://fox.wikis.com/wc.dll?Wiki~VisualFoxProDescription
[20] [Liu1993], pp. 23-25; p. 90
[21] weder in der Dissertation als auch nach Kenntnis der Autorin an anderer Stelle
[22] Colibri (COntext generation and LInguistic Tools for Bibliographic Retrieval Interfaces), s. Kap. 3.3 ff
[23] s.a. [Liu1993], pp. 66-85
[24] [Liu1993], pp. 6-17 (“Review of Literature”); pp. 233-235 („Appendix F. Selected Bibliography“)
[25] [Wajenberg1983], pp. 246-251. Arnold S. Wajenberg: “Marc Coding of DDC for Subject Retrieval”. Information Technology and Libraries, September 1983
[26] [DuncanWilliams1991] Elizabeth Duncan; James G. Williams: “A Rule Based System for Translating Dewey Decimal Numbers”. A Project Report. 1991
[27] [Liu1993], pp. 8-13. Ausführungen zu [Wajenberg1983] und [DuncanWilliams1991] basieren auf Liu1993.
[28] MARC-Felder „082“- bzw. „092“

Duncan & Williams erforschen die Möglichkeit, DDC-Notationen allgemeinverständlich (in DDC-Klassenbenennungen) automatisch zu übersetzen, was eine DDC-Notationszerlegung voraussetzt. Ihr auf die 20. DDC-Ausgabe stützendes Forschungsprojekt berücksichtigt die DDC-Klassen „330-339“ und „620-690“ der DDC-Haupttafel, für deren Übersetzung sie ca. 19 Regeln ermitteln. Nach Schätzung von Duncan & Williams[29] sind insgesamt ca. 30 Übersetzungregeln notwendig, um beliebige DDC-Notationen der 20. DDC-Ausgabe in Klassenbenennungen zu übersetzen.

Es folgen nun Details zu Liu's - auch (wie bei Duncan & Williams 1991) auf der 20. DDC-Ausgabe basierenden - DDC-Notationszerlegung, die er in sieben Schritten vornimmt und vorstellt[30]:

L1. Studium und Analyse der DDC-Regeln zur DDC-Notationssynthese
L2. Erstellung der Stichprobe
L3. Erstellung einer DDC-Wissensbasis
L4. Entwicklung des Computersystems „DND“
L5. Ausführung der DDC-Notationszerlegung
L6. Experten-Bewertung der Ergebnisse der DDC-Notationszerlegung
L7. Schlußfolgerungen aus den Ergebnissen für Anwendungsbereiche

Abb. 3.2.2 Liu's sieben Arbeitsschritte (hier L1-L7) zur DDC-Notationszerlegung

ad L1. DDC-Regeln[31]

Liu's Studium und Analyse der DDC-Hauptklasse „700“ mit ca. 2.600 Klassen[32] führen ihn auf 13 „Add“-Regeln, die er bei seiner DDC-Notationszerlegung einsetzt[33]:

```
AN 1: Add to base number X the numbers following Y in Z
AN 2: Add to base number X notation 001-999
AN 3: Add to base number X the numbers following Y in notation Z from T
AN 4: Add to base number X the numbers following Y in notation Z in
      [Internal] Table under T in Table U
AN 5: Add to base number X notation Y from Z
AN 6: Add to each subdivision identified by X as follows
AN 7: Add to each subdivision identified by X notation Y from Table under Z
AN 8: Add to each subdivision identified by X the numbers following Y in Z
AN 9: Add to X as instructed under Y
AN10: Add Standard Subdivisions as instructed under X
AN11: Add to base number X the numbers following Y in notation Z of this
      Table
AN12: Add the numbers following X in Y
AN13: Add 0 to the number from this Table and to the result add notation X
      from Table 2
```

Abb. 3.2.3 Zusammenstellung der von Liu klassifizierten 13 „Add“-Regeln (zu Klasse 700)

[29] laut [Liu1993], p. 12 allerdings ohne nähere Angaben und Begründung. Liu schreibt: „... One has to wonder about this estimate. ...“
[30] [Liu1993], pp. 19-27
[31] [Liu1993], p. 19; pp. 29-56
[32] [Liu1993], p. 19
[33] [Liu1993] pp. 28-39. „IV. Findings“. Liu merkt (S.33) an, dass sprachliche Varianten in den Anmerkungen vorkommen können, z.B. „Note: The phrase ‚base number' and ‚notation' may not occur.“

Weiterhin unterscheidet Liu zusätzlich zu AN1-AN13 vier in den Hilfstafel vorkommenden Fälle SS1-SS4 (SS: Standard Subdivisions)[34]:

```
SS1: The first digit of the remaining number after a match is zero, and the
     match number cannot be broken down using Standard Subdivisions
SS2: The first digit of the remaining number after a match is zero, and the
     match number can be broken down using Standard Subdivisions
SS3: The first digit of the remaining number after a match is not zero, and
     the match number has one of the following forms:
     X0X, X00.X, XX0.X, XX0.0X, XXX.0X, XXX.00X, XXX.000X, XXX.X0X,
     XXX.XX0X, XXX.XXX0X, or XXX.XXXX0X (X: any number)
SS4: The first digit of the remaining number after a match is not zero, and
     the match number is of the form XX0
```

Abb. 3.2.4 Liu's Klassifikation der mit Hilfstafel I erstellten Ziffernfolgen

Aus AN1-AN13 und SS1-SS4 erstellt Liu die 17 Regeln zur Zerlegung („rules for decomposition") A1-A13 und S1-S4:

```
A 1: If AN1, concatenate the number after "the numbers following" and the
     target number minus the base number[35]; insert a decimal after the third
     digit, if the concatenated number has more than three digits; match
     the number against the Schedule
A 2: If AN2, match the target number minus the base number against the
     Schedule. If the number to be matched has two digits or fewer, add
     zeros to the end of the number until a number with 3 digits is reached
A 3: If AN3, concatenate the number after "the number following" and the
     target number minus the base number; match this new number against the
     Table specified
A 4: If AN4, concatenate the number after "the numbers following" and the
     target number minus the base number; match this new number against the
     Internal Table specified
A 5: If AN5, match the target number minus the base number against the
     Table specified
A 6: If AN6, match the remaining number[36] against the Internal Table under
     the class where the note is located
A 7: If AN7, match the remaining number againt the Table specified
A 8: If AN8, concatenate the number after "the numbers following" and the
     remaining number; insert a decimal after the third digit if the
     concatenated number has more than three digits; match the number
     against the Schedule
A 9: Found only in Table 3 (1001-1009) and, therefore, not used
A10: If AN10, match the remaining number against Table 1
A11: If AN11, concatenate the number after "the numbers following" and
     the target number minus the base number; match this new number against
     the Table from which AN11 is obatined
A12: If AN12, concatenate the number after "the following" and the matched
     number plus remaining number; insert a decimal after the third digit,
     if the concatendated number has more than three digits; match the
     number against the Schedule
A13: If AN13, drop the leading zero of the remaining number and match the
     number against Table 2
```

Abb. 3.2.5 Liu's Regeln für Synthese-Anweisungen („Add notes")

[34] [Liu1993], p. 37-39

[35] [Liu1993], p. 41. Fußnote 15: „The target number refers to the complete number to be matched. The base number refers to the number immediately following the character string 'Add to [base number]' in a note. The target number minus the base number is the target number with the base number removed."

[36] [Liu1993], p. 45. Fußnote 17: „The remaining number refers to the number left after a match is made."

```
S1: If SS1, match the remaining number against Table 1[37]
S2: If SS2, concatenate the digits starting with the last zero of the match
    number and match the new number against Table 1
S3: If SS3, concatenate 0X (the last zero of the match number plus numbers
    after the zero, with decimal point removed) with the remaining number;
    match the new number against Table 1
S4: If SS4, concatenate 0 with the remaining number; match the new number
    against Table 1
```

Abb. 3.2.6 Liu's Regeln für Hilfstafel 1 (Standard subdivisions)

ad L2. Liu's Stichprobe[38]

Liu's Zufallsstichprobe stammt aus einer OCLC-Datenbank mit konstruierten DDC-Notationen. Den Umfang der Stichprobe bestimmt er nach der Formel („sample size for desired margin of error")[39]:

$$n = (2z^*/w)^2 p^*(1-p^*) \quad \text{(F3.1)}$$

mit

w : Fehlerbereich (margin of error), auch Genauigkeit

p^* : geschätzter Anteil zerlegbarer DDC-Notationen

z^* : tabellierte statistische Funktion

Für $p^* = 0.5$ vereinfacht sich die Formel (F3.1) zu

$$n = (z^*/w)^2 \quad \text{(F3.2)}$$

Bei einem Konfidenzniveau von 95% (entsprechend $z^* = 1.96$) und einem Fehlerbereich von $-0.04 \leq w \leq 0.04$ (Genauigkeit 4%) ergibt sich ein Stichprobenumfang von 600[40]. Die Stichprobe erstellt Liu folgendermaßen: von 11.662 bibliographischen bei einer Recherche gefundenen OCLC-Datensätzen der DDC-Klassen 700-799 werden von ihm nur die Datensätze in einer Datenbank abgespeichert, deren DDC-Notationen zerlegbar sind, in Bezug auf die Datenbank einmalig vorkommen, auf der 20., vollständigen DDC-Ausgabe beruhen und von der Library of Congress (LC) vergeben wurden. Von 1.743[41] DDC-Notationen (Grundgesamtheit), die diese Bedingungen erfüllen, ermittelt Liu 600 DDC-Notationen (repräsentative Stichprobe) per Zufall[42].

[37] [Liu1993], p. 53. Fußnote 18: „If the remaining number has more than one leading zeros, drop zeros until there is only one zero left."

[38] [Liu1993], pp. 20-23

[39] [MooreMcCabe1989], 3. Aufl.: [MooreMcCabe1996] p. 593. David S. Moore; George P. McCabe: "Introduction to the Practice of Statistics". 3rd ed. W. H. Freeman and Company, New York, 1996

[40] $n = (1.96/.08)^2 = 600$

[41] [Liu1993], p. 22. Es wurden alle 1.743 DDC-Notationen (und nicht nur die aus der Stichprobe stammenden) 600 DDC-Notationen zerlegt, da DND in der Lage ist, die Zerlegungsergebnisse in akzeptabler Antwortzeit bereitzustellen („the DND is capable of processing thousands of records in a couple of minutes")

[42] [Liu1993], p. 23. Die Zufallszahlen erzeugt Liu mit dem frei verfügbaren Statistikpaket EPISTAT

ad L3.-L4. Liu's Computersystem „Dewey Number Decomposer (DND)"[43]

Mit einem von Liu geschriebenen Konvertierungsprogramm werden die OCLC-Datensätze der Stichprobe aufbereitet und in einer Datenbank gespeichert. Jeder Datensatz dieser Datenbank enthält die zu zerlegende DDC-Notation, Buchtitel, Schlagwörter der „Library of Congress" (LCSH) und Datensatzidentifikation (record id). Werden eine/mehrere zu zerlegende DDC-Notation/en aus dieser Datenbank in DND eingegeben, gibt DND in Komponenten zerlegte Notationen („decomposed component numbers") aus. Laut Liu's Beschreibung besteht das interaktive DND-System aus den drei Hauptkomponenten Wissensbasis, Regelbasis und Inferenzmaschine. Außerdem verfügt das Computersystem DND über eine Erklärungskomponente[44], die die bei einer DDC-Notationszerlegung angewendeten Schritte und Regeln ausgibt. Die Wissensbasis (DDC database) besteht aus der DDC-Haupttafel 700, den Hilfstafeln 1-7[45] und den in Haupt- und Hilfstafeln enthaltenen internen Anhängetafeln (add tables). Eine Zusammenstellung von Liu's Regeln der Regelbasis sind den vorstehenden Abbildungen Abb. 3.2.5 und Abb. 3.2.6 zu entnehmen. Die Inferenzmaschine führt nach Eingabe von DDC-Notationen mithilfe der Wissens- und Regelbasis die DDC-Notationszerlegung durch. Werden mehrere zerlegbare DDC-Notationen eingegeben, generiert DND einen Bericht (report) und gibt eine Liste dieser DDC-Notationen mit Buchitel (associated book titles), Schlagwörtern (subject headings) samt ihrer Zerlegung in Komponenten aus. Zur Bewertung der DND-Zerlegungsergebnisse (Schritt L6, Abb. 3.2.2) erstellt Liu drei Listen mit je 200 DDC-Notationen für die drei DDC-Experten[46].

ad L5.-L6. Liu's Zerlegungsergebnisse mit Experten-Bewertung[47]

Im Anhang von Liu's Dissertation sind alle den drei DDC-Experten vorgelegten 600 DND-Zerlegungen vollständig dokumentiert. Die Experten kommen zu dem Schluß, daß Liu's Computersystem DND die 1.743 (s. Fußnote 41) DDC-Notationen

- 100% (nach Beseitigung eines trivialen Fehlers) korrekt zerlegt,
- 42 von Klassifizierern inkorrekt gebildete DDC-Notationen erkennt,
- 9 von Klassifizierern inkorrekt gebildete DDC-Notationen korrekt zerlegt.

Die in der Stichprobe enthaltenen von Klassifizierern inkorrekt gebildeten DDC-Notationen sind syntaktischer (42 DDC-Notationen) und semantischer (9 DDC-Notationen) Natur. Ob semantische DDC-Notationsbildungsfehler automatisch erkannt werden können, ist eine

[43] [Liu1993], pp. 19-20 (Overview of Procedures); pp. 23-25 (Instrumentation); pp. 90 (DND Systems Flowchart)

[44] [Liu1993], p. 25. „The DND has an explanation facility that can answer a ‚WHY' question by displaying the steps and rules by which a synthesized DDC number is decomposed. Library students learning DDC classification can benefit a great deal by studying such explanations for the decomposition."

[45] Bemerkung: Hilfstafel 7 ist in der 22. DDC-Ausgabe nicht mehr enthalten, die zuvor in Hilfstafel 7 enthaltene Information befindet sich ab der 22. DDC-Ausgabe in der Haupttafel und in der Hilfstafel 1 unter T1--08 (History and description with respect to kinds of persons). Vgl. [ChanMitchell2003], p. 12

[46] [Liu1993], p. 27. Die drei DDC-Experten zur Bewertung der Ergebnisse der DDC-Notationszerlegung sind: 1. Dr. Svenonius, library school at UCLA [Anmerkung d. Autorin: Prof. Elaine Svenonius ist Liu's (Doctor of Philosophy in Library and Information Science) Betreuerin und Gutachterin der Dissertation]; 2. Julie Beall, Assistant Editor, DDC Division at Library of Congress; 3. Bhagirathi Subrahmanyam, doctoral student, library school at UCLA

[47] [Liu1993], pp. 91-232. Appendixes C-E (Sample Decomposed Numbers 1-3) und pp. 56-64 (Results of the Decomposition)

offene Forschungsfrage und wird von Liu in Bezug auf DND kurz erörtert[48]. Die syntaktisch inkorrekt gebildeten DDC-Notationen unterteilt Liu in folgende Fälle[49]:

- fehlerhafte wiederholte Hilfstafel 1- Anwendung ohne Anweisung („errors due to more than one standard subdivision used“)[50]
- fehlende Ziffern („errors due to digits missing“)
- zu viel enthaltene Ziffern („errors due to extra digits used“)
- falsch benutzte Ziffer („errors due to wrong digit used“)
- an falscher Stelle gesetzter Dezimalpunkt („errors due to decimal point in wrong place“)
- Ziffernfolge in verdrehter Reihenfolge („errors due to digits in wrong order“)
- verkehrte DDC-Ausgabe („errors due to wrong edition used“)

ad L7. Schlußfolgerungen und Anwendungsbereiche[51]

Liu' Schlußfolgerungen aus seinen Untersuchungen sind bereits am Anfang des Kapitels 3 (Abb. 3.2.1) wiedergegeben. Nachdem Liu die automatische Zerlegbarkeit für eine repräsentative Stichprobe von zerlegbaren DDC-Notationen der Klasse „700“ nachgewiesen hat, erörtert er einige nützliche Anwendungsmöglichkeiten am Ende seiner Arbeit:

- Information Retrieval und automatische Indexierung: Verbesserung der Retrievalergebnisse (precision, recall) durch beliebige Kombination (facets) der DDC-Teilnotationen (component numbers) als Suchbegriffe

- Entwurf und Entwicklung von Expertensystemen für die (semi-)automatische DDC-Notationsanalyse und DDC-Notationssynthese

- Automatische Indexierung: Bereicherung der Erschliessung bibliographischer Datensätze durch automatische Hinzufügung aller DDC-Teilnotationen einer DDC-Notation

- Sprachkonvertierung (switching language development[52]): Durch die in DDC-Notationen vorliegende künstliche Zwischensprache besteht - unter Nutzung des in ca. 30 Sprachen vorliegenden DDC-Systems - die Möglichkeit, Recherchen mit multilingualen Deskriptoren durchzuführen

- DDC-Lehrsystem: Mithilfe einer Erklärungskomponente eines Expertensystems können die bei der DDC-Notationsbildung Schritte und Regeln gelernt werden

- Qualitätskontrolle der DDC-Notationsvergabe: Zur Klassifizierungs-Unterstützung und Prüfung der Korrektheit vergebener DDC-Notationen.

[48] [Liu1993], pp. 61-64. „C.3 Analysis of Errors That Cannot Be Identified“
[49] [Liu1993], pp. 58-61
[50] [http://connexion.oclc.org/] „WebDewey Help – Introduction”, Chap. 8.6 (Number Building): „Do not add multiple standard subdivisions to the same number except when specifically instructed to do so ...”
[51] [Liu1993], pp. 66-85
[52] [Liu1993], pp. 80, Fußnote 25: „A switching language is a device that can be used to convert from one vocabulary to another.”

3.3 VZG-Colibri-Ansatz

3.3.1 Problemstellung

Hauptziel des VZG-Projektes Colibri ist eine verbesserte (einheitliche) Inhaltserschliessung zur fachgebietsbezogenen Suche auf Basis der Dewey-Dezimalklassifikation. An anderer Stelle[53] sind diesbezüglich folgende Forschungsfragen formuliert worden:

Q1-COLIBRI:
„Ist es möglich, eine inhaltlich stimmige DDC-Titelklassifikation aller GVK[54]-PLUS-Titel automatisch zu erzielen?“

Q2-COLIBRI:
„Ist es möglich, molekulare DDC-Notationen automatisch (eindeutig) zu zerlegen?“[55]

Q3-COLIBRI:
„Ist es möglich, mit Hilfe atomarer DDC-Notationen die Klassifikation und Suche zu verbessern?“

Die in diesem Bericht im Mittelpunkt stehenden Untersuchungen zu „Q2-COLIBRI“ stellen einen wichtigen Meilenstein für die beiden anderen Fragestellungen „Q1-COLIBRI“ und „Q3-COLIBRI“ dar. Wie im vorangegangen Kapitel erwähnt, bilden Liu’s Ergebnisse den Ausgangspunkt, von dem aus - wie nachfolgend beschrieben - auch neue Wege beschritten werden. Vor Darstellung der „Q2-COLIBRI“-Details, wird auf die Einbettung der DDC-Notationsanalyse-Komponente (vc_day) in das DDC-Suchsystem (vc_ds) eingegangen.

3.3.2 DDC-Suchsystem (vc_ds) mit Analysekomponente (vc_day)

Gegeben sei ein Titeldatenbestand T mit dem GVK-Titeldatensatz „ppn 319238385“. Der Titeldatensatz ist z. Zt. mit folgenden inhaltstragenden Elementen erschlossen:

Autor	{Jim Jermyn}
Titel	{Himalaya garden – growing plants from the roof of the world}
LCSH[56]	{Alpine gardens, Alpine garden plants, Alpine gardens / Himalya Mountains, Alpine garden plants / Himalya Mountains, Native plants for cultivation / Himalya Mountains}
Einzelschlagwort:	{Himalaja, Gebirgspflanzen}
LCC[57]	{SB459}
DDC	{635.9528095496}
Basisklassifikation:	{42.44 Pflanzengeographie}

[53] [Reiner2003] Ulrike Reiner: „VZG-Projekt Colibri - Überblick, Stand, Ergebnisse - Juli-Dezember 2003“, VZG-Colibri-Bericht 1/2003 und
[Reiner2004] Ulrike Reiner: „VZG-Projekt Colibri / DDC (Präsentation 4. FAG EI) http://134.28.50.10/mambo/downloads/colibri_fagei-04-09-03.pdf
[54] GVK-PLUS: Gemeinsamer VerbundKatalog mit Online Contents (http://gso.gbv.de/)
[55] Zerlegung in atomare (semantisch sinnvolle) DDC-Notationen, genauer in Kap. 3.3.2
[56] LCSH: Library of Congress Subject Headings
[57] LCC: Library of Congress Classification

Exkurs1: Colibri-Modell [Anfang]

Im Colibri-Modell[58] werden u.a. Deskriptor, Deskriptorwert, Titeldatensatz, DDC-Klasse und DDC-Basis definiert:

Deskriptor
Jedes inhaltsbeschreibende Element der PICA3(PICA+)-Kategorien[59], wobei folgende Deskriptoren unterschieden werden:

3000(028A), 300x(028B), 301x(028C), 4000(021A), 4222(046M), 4241(039B), 5010(045F), 502x(045B), 5030(045A), 5040(045C), 5050(045E), 5075(045T), 5080(045U), 5100(041A), 51xx(041A), 530x(045Q), 531x(045R), 5500(044A), 5510(044C), 5520(044E), 5530(044F), 5540(044G), 555x(044K), 558x(044L), 60xx(145Z), 65xx(144Z), 67xx(245Z), 68xx(244Z)

Deskriptorwert
Wert einer PICA-Kategorie

Titeldatensatz (T)
Menge von Deskriptorwerten

DDC-Klasse (D)
Menge von Deskriptorwerten

DDC-Basis
Menge von DDC-Klassen

Exkurs1: Colibri-Modell [Ende]

Daraus ergeben sich als DDC-Basis und DDC-Klasse:

DDC-Basis := {635.9528095496}

DDC-Klasse 635.9528095496 :=
{Alpine_garden_plants, Alpine_garden_plants#Himalaya_Mountains, Alpine_gardens,
Alpine_gardens#Himalaya_Moutains, garden, growing, Himalayan, Jim#Jermyn,
Native_plants_for_cultivation#Himalaya_Mountains, plants, roof, SB459, world}

Wird die Anfrage „635.9528095496“ an das (zu entwerfende) DDC-Suchsystem[60] gestellt, liefert es die Antwortmenge „ppn 319238385“. In der Anfragesprache IQL[61] fomuliert:

$$I((\lambda x)\ ppn\_dno(x,635.9528095496)) = \{ppn\,319238385\}$$

d.h.: die Bedeutung der Anfrage ist „ppn 319238385“.

[58] [Reiner2004], S. 11-14
[59] Bedeutung der PICA-Kategorien z. B. in http://www.gbv.de/du/katricht/inhalt.shtml
[60] Boolesche Suche oder Freitext-Suche, s. z. B. [Reiner 1991] Reiner, Ulrike: „Anfragesprachen für Informationssysteme. Reihe Informationswissenschaft der Deutschen Gesellschaft für Dokumentation (hrsg. von W. Neubauer, DGD). Frankfurt am Main, 1991
[61] [Reiner1991], S. 79ff

Exkurs2: Interpretation von Anfragen [Anfang]

Durch eine Interpretation I - als Abbildung definiert - wird jeder Anfrage aus der Menge aller Anfragen genau eine Teilmenge aus dem Titeldatenbestand T, hier aus einem Titeldatensatz bestehend, zugeordnet. In der universellen Anfragesprache IQL werden Anfragen in der Sprache der Prädikatenlogik 1. Stufe formuliert, wobei λ „die Menge, welche" bedeutet, x steht für eine Individuenvariable. ppn_dno (mit der Bedeutung: „ ppn ist mit der DDC-Notation dno klassifiziert") ist eine 2-stellige Prädikatenkonstante und 635.9528095496 eine Individuenkonstante.

Exkurs2: Interpretation von Anfragen [Ende]

Ein DDC-Suchsystem[62] ohne DDC-Notations-Analysekomponente liefert (bei gegebenem T) auf die Anfrage $Horticulture$ (Gartenbaukunst) die leere Antwortmenge:

$$I((\lambda x)\ ppn\_cap(x, Horticulture)) = \{\}$$

Wird die DDC-Notation 635.9528095496 jedoch einer DDC-Notationsanalyse unterzogen und der Titeldatensatz $ppn319238385$ mit der erhaltenen Analyseinformation angereichert, kann das System auf die Anfrage als Antwort den einzigen in T enthaltenen Titeldatensatz ausgeben (vgl. „Q3-COLIBRI"):

$$I((\lambda x)\ \ ppn\_cap(x, Horticulture)) = \{ppn319238385\}$$

Eine (vollständige[63]) DDC-Notationsanalyse mit der Analysekomponente „***vc_day***" (**v**zg **c**olibri_**d**dc notation **a**nal**y**zer, s. Abb. 3.3.2.3) ermittelt für den Titeldatensatz mit der DDC-Notation 635.9528095496 folgende zusätzliche DDC-Klassenbenennungen (ppn_cap), die zur Indexierung verwendet werden können (Detaillierte Ausgabe in Abb. 4.10):

```
{Agriculture, Alpine_plants, Asia____Orient____Far_East,
 Flowers_and_ornamental_plants, Garden_crops_(Horticulture)____Vegetables,
 Groupings_by_climatic_factors, Groupings_by_environmental_factors,
 Historical,_geographic,_persons_treatment, Nepal, Other_jurisdictions,
 South_Asia____India, Technology}
```

Abb. 3.3.2.1 Die in DDC-Notation 635.9528095496 enthaltenen Klassenbenennungen

Im Weiteren nennen wir die von DDC-Expert(inn)en nach Regeln zusammengesetzte DDC-Notationen (wie 635.9528095496) *molekulare DDC-Notationen* (oder *molekulare DDC-Nummern*) und ihre einzelnen DDC-Notationsteile *atomare DDC-Notationen* (oder *atomare DDC-Nummern*).

Nachdem mit vorstehenden Ausführungen die Forschungsfrage „Q2-COLIBRI" erläutert wurde, wird das DDC-Suchsystem (***vc_ds*** : **v**zg **c**olibri_**d**dc **s**earch system) schematisch in einem Modell (Abb. 3.3.2.2) mit möglichen Anfragen / Antworten (auch Abb. 3.3.2.4) abgebildet. Die Namen der Teilkomponenten des DDC-Suchsystems ***vc_ds*** sind kursiv und falls mit der Realiserung / Implementierung begonnen wurde fett gedruckt (Abb. 3.3.2.3).

[62] vgl. Fußnote 60
[63] bei feinster Zerlegungs-Granularität der DDC-Notationen „635.9528095496" in atomare DDC-Notationen

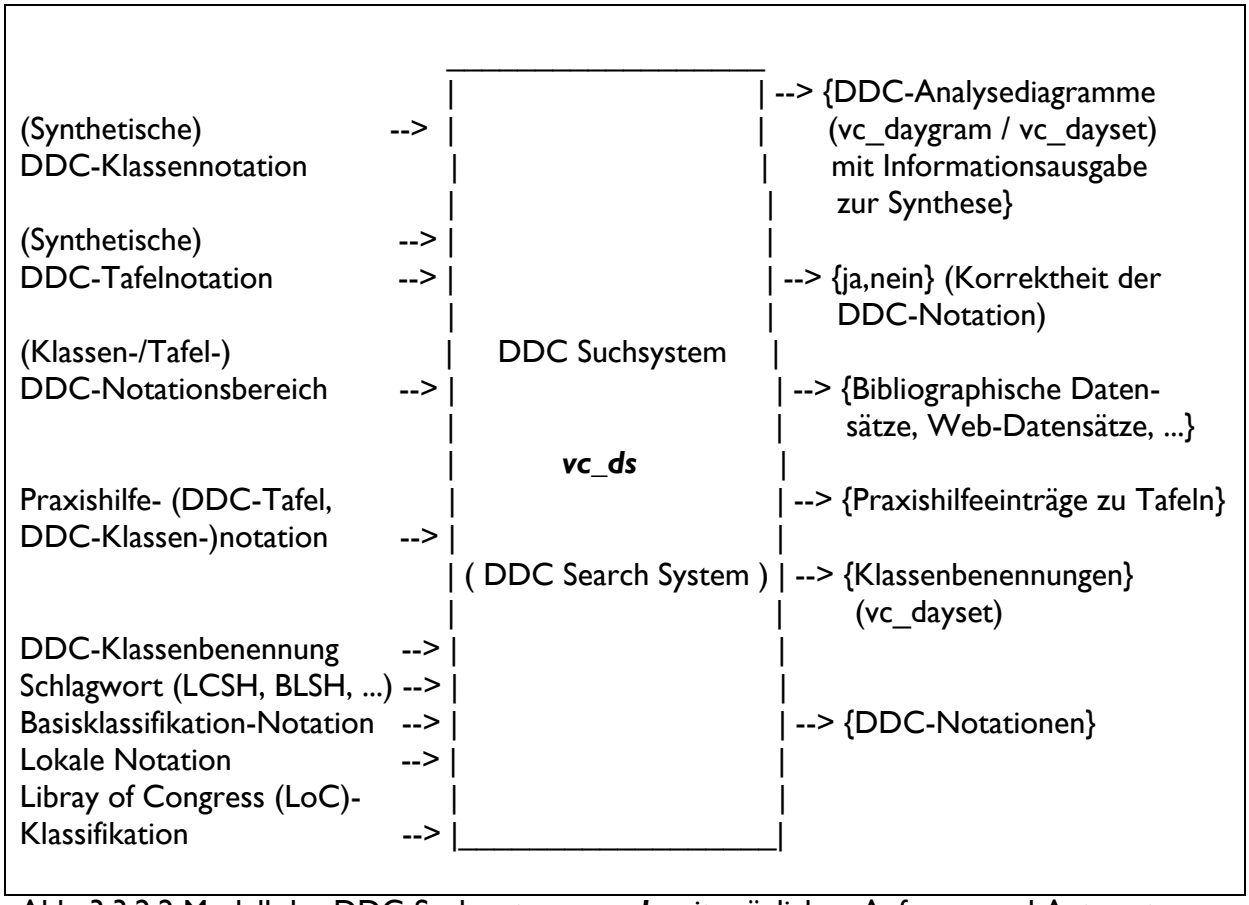

Abb. 3.3.2.2 Modell des DDC-Suchsystems ***vc_ds*** mit möglichen Anfragen und Antworten

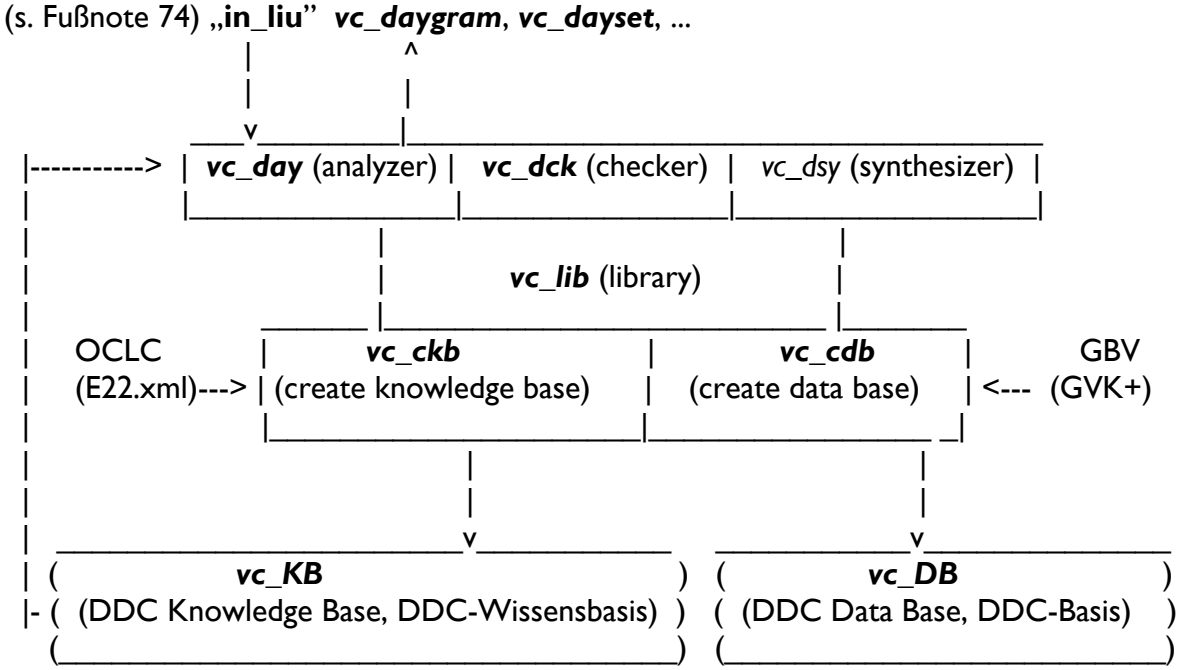

Abb. 3.3.2.3 DDC-Suchsystem ***vc_ds*** mit Systemkomponenten

Das DDC-Suchsystem ***vc_ds*** sucht z. B.

a) nach Eingabe von DDC-Notationen

1. eine DDC-Klassenbenennung
2. alle in ihr enthaltenen DDC-Klassenbenennungen oder DDC-Notationsteile (vc_dayset)
3. Identifikationsnummern (z. B. ppn) oder Anzahl bibliographischer Datensätze, die diese DDC-Notationen enthalten

b) (atomare, molekulare) DDC-Notationen nach Eingabe von

1. Klassenbenennungen (captions)
2. Schlagwörtern (z. B. aus DDC relative index, KWIC)
3. LCSH (Library of Congress Subject Headings, 044A [PICA+])
4. Basisklassifikation-Notation (BK, 045Q [PICA+])
5. British Library Subject Headings (BLSH, 044G [PICA+])
6. Einzelschlagworte (044K [PICA+])
7. Lokale Notationen (144Z, 244Z [PICA+])
8. Library of Congress Classification (LCC) (045A [PICA+])
9. Anmerkungen (notes)

Es folgen Beispiele zu a1.-b9. zur Anwendung des DDC-Suchsystems:

```
a1.) 635.9528        -> {Alpine plants}
a2.) 635.9528        -> {Agriculture, Alpine plants, Flowers and ornamental
                         plants, Garden crops (Horticulture) Vegetables,
                         Groupings by climatic factors, Technology} oder
                        {6,63,635,635.9,635.95,635.952,635.9528}
a3.) 635.9528        -> {ppn0237659972, ppn0268321213, ppn0315497610,
                         ppn0323095682, ppn0353462993} bzw.
     635.9528        -> |5| (Anzahl 5: d.h. es sind 5 GVK-Titeldatensätze
                         mit der DDC-Notation „635.9528“ klassifiziert)
b1.) Alpine plants   -> {635.9528, 581.7538}
b2.) Alpine          -> {T2--79443, 577.538, B578.7538, B581.7538,
                         635.9528, 796.935}
b3.) Alpine gardens  -> 635.9528095496
     Alpine garden plants -> 635.9528095496
     Alpine gardens_Himalaya Mountains -> 635.9528095496
     Alpine garden plants_Himalaya Mountains -> 635.9528095496
     Native plants for cultivation_Himalaya Mountains -> 635.9528095496
     Alpine garden plants_West (U.S.) -> 635.95280978
     Native plants for cultivation_West (U.S.) -> 635.95280978
b4.) 42.44 (Pflanzengeographie) -> 635.9528
b5.) Alpine garden plants -> 635.9528
     Rock gardens    -> 635.9528
     Rock plants     -> 635.9528
     Alpine gardens  -> 635.9528
b6.) Gebirgspflanzen -> 635.9528
b7.) bio 453.2       -> 635.9528
b8.) SB421           -> 635.95280978
b9.) in Anmerkungen „... Alpine plants ...“ -> 635.9528
```

Abb. 3.3.2.4 Beispielanfragen und -antworten (mit WebDewey/GVK[64] ermittelt)

[64] http://connexion.oclc.org/ bzw. http://gso.gbv.de/

Das (aufgrund der Modularität einfach erweiterbare) DDC Suchsystem ***vc_ds*** besteht - wie oben abgebildet - aus folgenden Komponenten:

```
Hauptkomponente (ruft andere (autonome) Komponenten je nach Option auf):
vc_ds: vzg colibri – ddc search system

autonome Komponenten
vc_cdb: vzg colibri – create ddc data base
vc_ckb: vzg colibri – create ddc knowledge base
vc_dck: vzg colibri – ddc checker
vc_day: vzg colibri – ddc number analyzer
vc_dsy: vzg colibri – ddc number synthesizer
vc_lib: vzg colibri - library
```

Der Prototyp ***vc_ds*** ist in der Sprache „gawk“[65] unter Linux realisiert, je nach Aufrufoption (Idee skizziert, noch nicht realisiert) wird die entsprechende Systemantwort ausgegeben:

NAME

vc_ds - search for (atomar, molecular), e. g., DDC numbers, vc_daygram (DDC analysis diagram), vc_dayset (DDC analysis result set), DDC captions, DDC Relative Index terms, LCSH terms, BLSH terms, controlled term, local classification (e. g. Basisklassifikation), DDC records in a given DDC data base

SYNOPSIS

vc_ds [OPTIONS]

DESCRIPTION

Dependent on the given input, the VZG Colibri DDC Search System vc_ds either searches for DDC (Dewey Decimal Classification) numbers or outputs DDC numbers. DDC numbers may be atomar or molecular.

OPTIONS

- d data base 1,[data base 2, data base 3, ...]
- k knowledge base 1,[knowledge base 2, knowledge base 3, ...]
- h help
- i [in_no | input-file] (in_no: input notation)
- o [output-file]
- p explain
- v version

Abb. 3.3.2.5 Aufruf und Optionen des Kommandos „***vc_ds***“

```
Beispielaufruf
vc_ds –d vc_DB –k vc_KB –i 635.9528095496 –o vc_daygram_635.9528095496 –p
explain
```

[65] [Herold1999a] Helmut Herold: „Linux-Unix-Profitools. awk, sed, lex, yacc und make”. 3. überarb. Aufl., Addison-Wesley, Bonn u.a., 1999 (http://cm.bell-labs.com/cm/cs/awkbook/)
s. a. Linux “man page” zu “gawk” “NAME gawk - pattern scanning and processing language ... DESCRIPTION Gawk is the GNU Project's implementation of the AWK programming language ... ” und Test-Entwicklungsumgebung zu „Q1-COLIBRI“ in [Reiner2003], S. 7ff

Für die Ermittlung der Systemantwort zum angegebenen Beispiel eines Aufrufs des DDC-Suchsystems werden die Systemkomponenten ***vc_DB***, ***vc_KB***, ***vc_day***, ***vc_dck*** und ***vc_lib*** benötigt. Die Softwarebibliothek ***vc_lib*** enthält von allen Teilkomponenten des DDC-Suchsystems gemeinsam verwendete benutzerdefinierte (im Unterschied zu built-in) gawk-Funktionen. Die Komponente ***vc_dck*** prüft die syntaktische Korrektheit von DDC-Notationen (oder DDC-Klassenbenennungen). Die DDC-Basis ***vc_DB*** bzw. DDC-Wissensbasis ***vc_KB*** werden durch die Komponenten ***vc_cdb*** bzw. ***vc_ckb*** erstellt, wobei als Eingabedateien GVK-PLUS-Datensätze (im PICA+ Format) bzw. OCLC-DDC-Datensätze (im XML-Format) dienen. Die für diesen Bericht wichtigste Teilkomponente ***vc_day*** erstellt als Hauptaufgabe DDC-Analysediagramme („***vc_daygram***“)[66] aus molekularen DDC-Notationen. Die Erstellung der DDC-Basis ***vc_DB*** und deren Details sind Inhalt eines anderen Berichts[67], weshalb hier nicht weiter darauf eingegangen wird. Anders verhält es sich mit der DDC-Wissensbasis ***vc_KB***: sie ist wesentlicher Bestandteil einer DDC-Notationsanalyse (und auch DDC-Notationssynthese). Wie oben erwähnt, verwendet die Komponente ***vc_ckb*** zur Erstellung der Wissensbasis ***vc_KB*** das (elektronische) DDC-System (22. Ausgabe) im XML-Format (Datei E22.xml). Es folgt zunächst eine Charakterisierung des DDC-Systems für die Aufgabe der DDC-Notationsanalyse.

Das DDC-System ist ein intellektuell erstelltes komplexes Klassifikationswerk, an dem seit über in einem Jahrhundert gearbeitet wird. Es besteht aus Haupttafeln, Hilfstafeln und anderen Tafeln, Regeln, Instruktionen, Begriffen, synthetischen Notationen und Klassenbenennungen. Das darin enthaltene DDC-Expertenwissen ist in der gedruckten und der elektronischen (online) Version an unterschiedlichen Stellen niedergelegt. Die Vorschriften zur Notationsbildung sind jeweils an den entsprechenden Systemstellen (meist mit zahlreichen weiteren An- und Verweisungen) dokumentiert und für Klassifizierer(innen) gedacht, die Schritt für Schritt DDC-Notationen erstellen. Die verteilte Dokumentation erschwert jedoch den Gesamtüberblick und ist für eine direkte automatische Verarbeitung[68] ungeeignet. Deshalb gehört zur Beantwortung der Forschungsfrage „Q2-COLIBRI“ u. a. sowohl das Erarbeiten einer geeigneten Repräsentation des vorliegenden DDC-Systems als auch das Auffinden eines (effizienten) Analyse-Algorithmus.

Wegen des umfangreichen im DDC-System enthaltenen Expertenwissens ist es naheliegend[69], auf Methoden aus dem Bereich der Künstlichen Intelligenz zurückzugreifen. Ein wesentlicher Teil eines Problems ist gelöst, wenn eine gute Repräsentation (kleiner Zustandsraum, wenige Regeln) gefunden ist:

„To solve a problem using a production system, we must specify the global database, the rules, and the control stratey. Transforming a problem statement into these three components of a production system is often called the *representation problem* in AI. Usually there are several ways to so represent a problem. Selecting a good representation is one of the important arts involved in applying AI techniques to practical problems. ... The process required to represent problems initially and to improve given representations are still poorly

[66] Beispiele in Kapitel 4
[67] [Reiner2005] Ulrike Reiner: „DDC-Basis: Grundlage der automatischen DDC-Klassifikation“, VZG-Colibri-Bericht 1/2004 [in Arbeit]
[68] d. h. Algorithmisierung des verteilten, natürlichsprachlichen Textes
[69] so auch [Liu1993]

understood. It seems that desirable shifts in a problem's representation depend on experience gained in attempts to solve it in a given representation. This experience allows us to recognize the occurrence of simplifying notions, such as symmetries, or useful sequences of rules that ought to be combined into macro-rules."[70]

Viele Probleme können mit Produktionssystemen aus dem Bereich der Künstlichen Intelligenz (KI) gelöst werden[71], z. B. das „Problem des Handlungsreisenden" (Operations Research), symbolische Integration (Mathematik), Analyse und Generation von (Kandidaten-) Strukturen organischer Verbindungen[72] (Chemie) und Syntaxanalyse natürlicher Sprache (Linguistik).

Auch wenn das DDC-System nicht die Komplexität eines KI-Systems aufweist, kann sich die Betrachtungsweise als KI-Produktionssystem auf Modellebene als hilfreich erweisen. Ein KI-Produktionssystem besteht aus den drei im o. g. Zitat genannten Hauptkomponenten: globale Datenbank, Menge von Produktionsregeln und Kontrollstrategie. Die „globale Datenbank" ist die Datenstruktur, bspw. eine einfache Nummernmatrix, aber z. B. auch eine sehr grosse Indexdatei. Die globale Datenbank wird durch anwendbare[73] Produktionsregeln verändert. Die Kontrollstrategie wählt - aus ggf. mehreren anwendbaren Regeln - eine Regel aus und beendet die Berechnung, wenn die Schlussbedingung der globalen Datenbank erfüllt ist. Eine Lösung des Problems der DDC-Notationsanalyse mit einem Produktionssystem könnte so aussehen:

Globale Datenbasis

Sie enthält am Anfang eine molekulare DDC-Notation, dann molekulare und atomare DDC-Notationen bis zur Erfüllung einer Terminationsbedingung.

Produktionsregeln

Sie werden aus DDC-Anweisungen gebildet.

Kontrollsystem

Es steuert die Auswahl und die Anwendung der Regeln.

Mit dem Produktionssystem kann dann die Frage beantwortet werden, ob eine beliebige Zeichenkette eine DDC-Notation einer DDC-Sprache ist (Parsing-Problem).

Zur Vorbereitung einer (zu entwerfenden) formalen vc_DDC-Sprache folgt ein Syntaxbaum für die 32. DDC-Notation aus der von Liu gezogenen Stichprobe (in_liu_32_to_analyze[74]):

[70] [Nilsson1990], p. 19; p. 29 Nils J. Nilsson: Principles of Artificial Intelligence. Tioaga Publishing Co., Palo Alto, 1980

[71] a.a.O., pp. 27ff

[72] realisiert im erfolgreichen Expertensystem DENDRAL

[73] Eine Regel ist anwendbar, wenn ihre Vorbedingungen von der globalen Datenbank erfüllt sind.

[74] Die in [Liu1993], pp. 91-114 „Appendixes C., D. und E." wiedergegebenen DDC-Notationen („Sample Decomposed Numbers 1, 2 3") wurden von der Autorin durchnummeriert (in_liu_1, ..., in_liu_600) und als Test- und Eingabedatei „in_liu" für die Analysekomponente ***vc_day.awk*** manuell erstellt

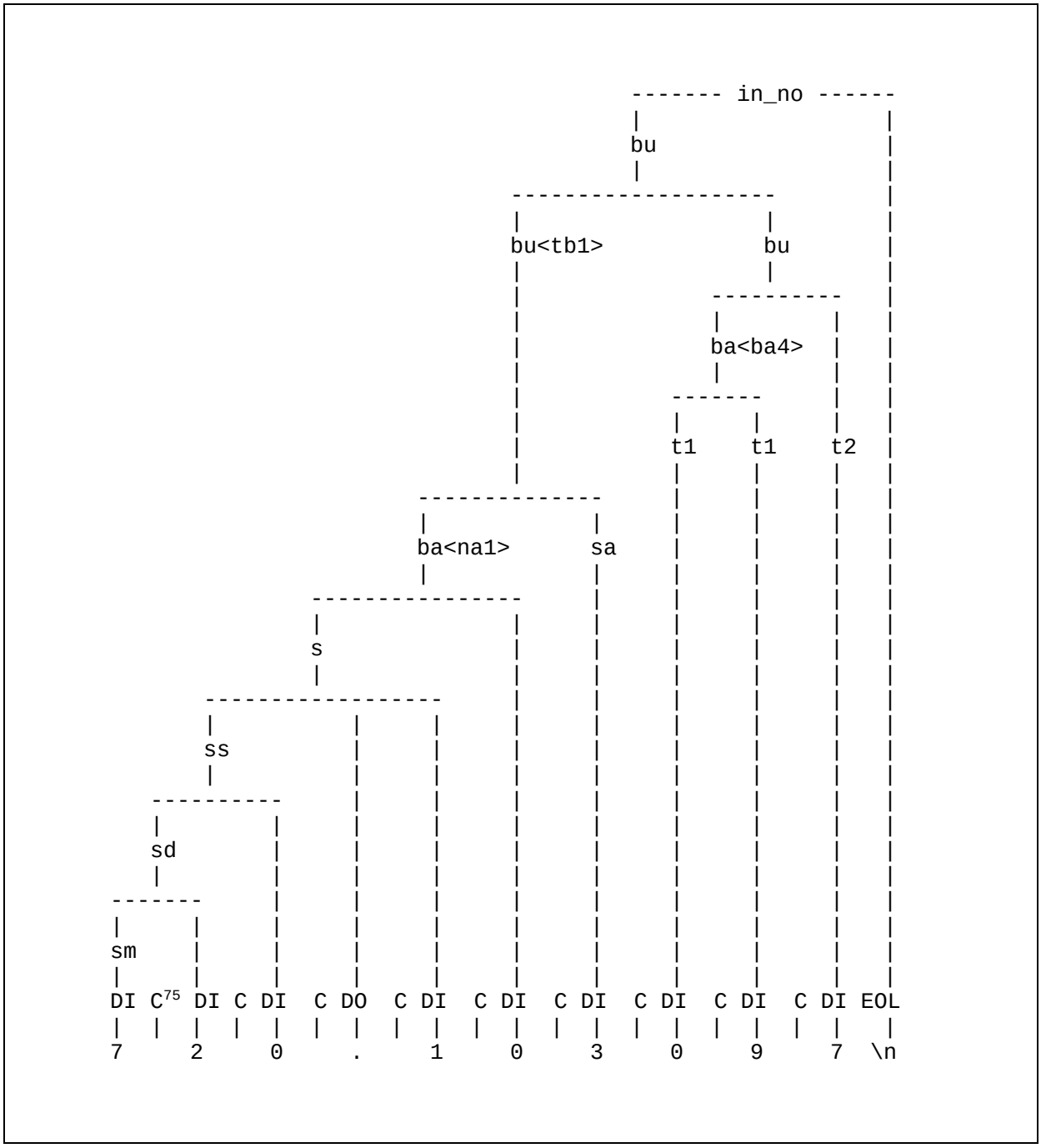

Abb. 3.3.2.6 Syntaxbaum (parse tree) für die DDC-Notation “720.103097“ <in_liu_32_to_analyze>

In Abb. 3.3.2.7 sind Erläuterungen zur vc_DDC-Metasprache zusammengestellt:

[75] „DI“ (digit) steht für Ziffer; „C“ (concatenation operator) für Operator zur Ziffern(Zeichen)verkettung; „Die *Konkatenation* zweier Zeichenketten ist die Zeichenkette, die sich ergibt, wenn auf die erste folgend direkt – ohne Zwischenraum – die zweite Zeichenkette dahinter geschrieben wird.“ In: [HopcroftUllman1992] John E. Hopcroft; Jeffrey Ullman: „Einführung in die Automatentheorie, formale Sprachen und Komplexitätstheorie.“ 1. Nachdruck, Addison-Wesley Publishing Company, Bonn u.a., 1992, S. 1

Legende vc_DDC-Metasprache (Kurzform)	Legende vc_DDC-Metasprache(Langform) / Erläuterungen mit OCLC Terminologie (vgl. Abb. 2.2, Abb. 3.3.3.16)
C	Concatenation operator
DI	DIGIT
DO	DOT (OCLC: Dewey decimal point)
EOL	End-Of-Line
ba	bano (OCLC: base number)
bu	buino (OCLC: built class number)
s	schedno (OCLC: schedule number)
sa	schedno_atom
sm	schedno_main
sd	schedno_division
ss	schedno_section
t	tabno (OCLC: table number)
ta	tabno_atom
t1,t2, ..., t6	tabno (t or ta) of table1,...of table2,...,of table6 (OCLC notation : T1--, T2--, ..., T6--)
<tb1>	apply table 1
<ba1>, <ba3>, <ba4>,	apply “rule” “ba1”, “ba3”, “ba4” (OCLC-XML tags)
<ba5>, <na1>, ...,	apply “rule” “ba5”, “na1”, ..., (OCLC-XML tags)
<na6>, <nfa>	apply “rule” “na6”, “nfa” (OCLC-XML tags)

Legende vc_DDC-Metasprache (Langform)	Erläuterungen aus OCLC-WebDewey (Glossary)
bano	Base number (“Table T3A. Subdivisions for Works by or about Individual Authos. Notes: ... Look in the schedule 810-890 to find the base number for the language. The base number may be identified in an add note, e.g., at 820.1-828 (“add to base number 82”) or another note, e.g., at 896 (“896.392 Swahili”); otherwise it is the number given for the literature, e.g., Dutch-language literature 839.31
buino	Built number (a number constructed according to add instructions stated or implied in the schedules or tables)
clano	Class number (notation that designates the class to which a given item belongs)
dno	DDC number
dno_atom	DDC number, part of a number (with meaning)
dno_span	Span of numbers (number span, range of numbers)
hookno	Hook number (a number in the DDC without meaning in itself; but used to introduce examples of the topic. Headings begin with 'Miscellaneous', 'Other', 'Specific')
in_no	input notation (vc_day)
preno	the numbers following <preno> in <span of numbers> (vc_KB)
schedno	schedule number
schedno_main	schedule number (main)
schedno_div	schedule number (divisions)
schedno_sec	schedule number (sections)
schedno_add	schedule number with add_table (<ba1>,<ba3>,<ba4>,<ba5>)
schedno_atom	part of a schedule number
schedno_id	designated schedule number (identified by a symbol)
schedno_nc	schedno_non_caption [= unassigned dno]
tabno	table number
tabno_add	table number with add_table (<ba1>,<ba3>,<ba4>,<ba5>)
tabno_atom	part of a table number

Abb. 3.3.2.7 Legende der vc_DDC-Metasprache

In Weiterführung der Überlegungen zu einer formalen DDC-Sprache ergibt eine Analyse des DDC-Systems (aus im Bericht angegebenen Literaturquellen) folgenden Zusammenhang für DDC-Notation(steil)en.

Eine DDC-Klassennotation (*clano*) ist entweder eine Haupt- (*schedno*) oder eine Hilfs-Tafelnotation (*tabno*) und diese ist entweder eine (einzelne) DDC-Klassennotation (*dno*) oder ein DDC-Klassennotationsbereich (*dno_span*). Ein *dno_span* besteht aus zwei mit einem Bindestrich verbundenen (einzelnen) *dno*'s: *dno_span1* und *dno_span2*. Eine *dno* ist entweder zerlegbar, d.h. molekular (*dno_mol*) oder nicht weiter zerlgebar, d.h. atomar (*dno_atom*). *dno_mol* und *buino* sind unterschiedliche Bezeichnungen für eine zusammengesetzte (synthetische) DDC-Klassennotation. Eine *dno_mol* ist entweder 1. eine *dno_mol* gefolgt von einer *dno_atom* oder 2. eine *schedno_atom* oder *schedno_mol* gefolgt von einer (atomaren oder molekularen) DDC-Notation aus Hilfstafel 1 (*t1_atom* oder *t1_mol*) oder 3. eine Grundnotation (*bano*) gefolgt von einer *dno_atom* oder *dno_mol.* Eine *bano* ist entweder eine *schedno* oder eine *tabno*. Eine *dno_atom* der Haupttafel (*schedno_atom*) oder einer Hilfstafel (*tabno_atom*) besteht aus soviel verketteten Ziffern, deren Interpretation eine semantische Einheit ergibt. Veranschaulicht:

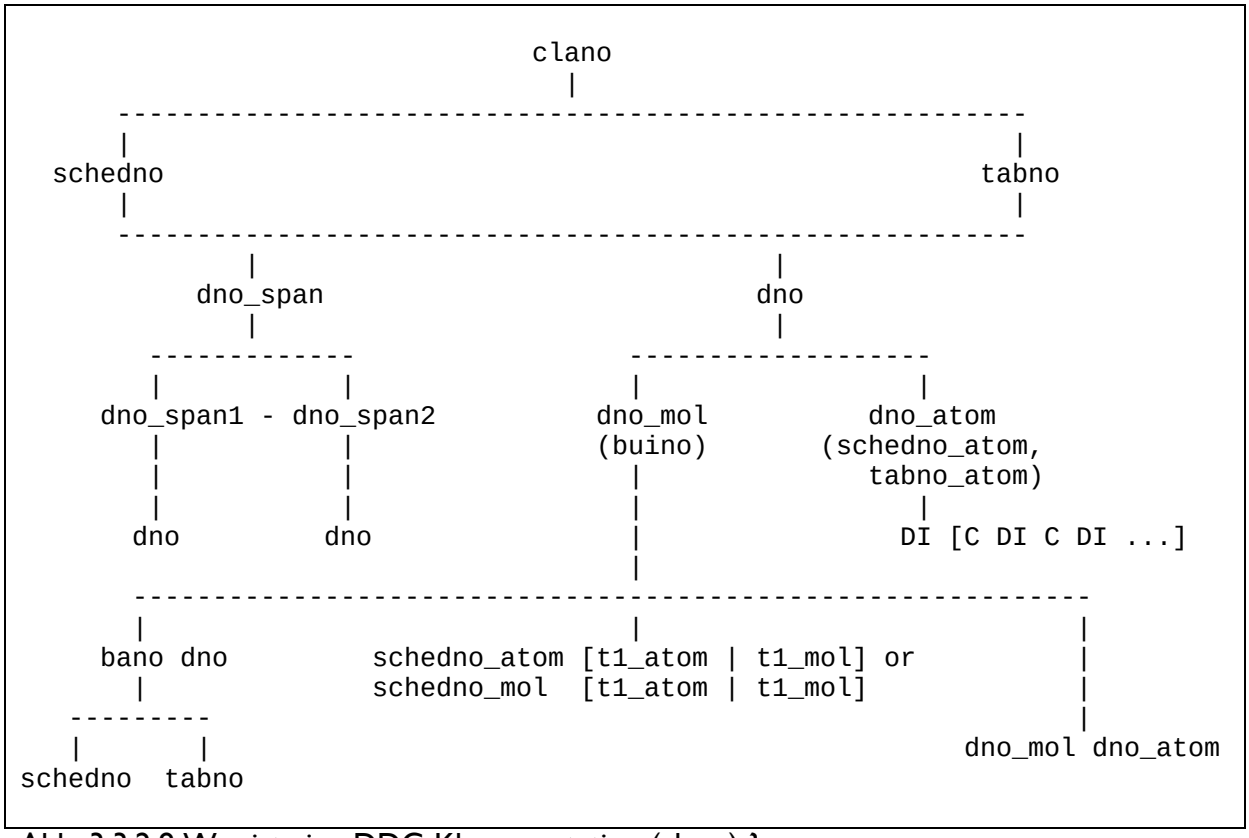

Abb. 3.3.2.8 Was ist eine DDC-Klassennotation (clano) ?

Nach den vorstehenden Überlegungen wird eine vc_DDC-Grammatik („rewrite rules") und eine vc_DDC-EBNF[76] erstellt:

[76] vgl. z. B. http://www.garshol.priv.no/download/text/bnf.html; Formal grammar of XML given using a simple Extended Backus-Naur Form (EBNF) notation: http://www.w3.org/TR/WD-xml-961114.html

```
schedno                     -> clano
tabno                       -> clano
dno                         -> schedno
dno                         -> tabno
dno_span                    -> schedno
dno_span                    -> tabno
dno_span1-dno_span2         -> dno_span
dno                         -> dno_span1
dno                         -> dno_span2
dno_atom                    -> dno
dno_mol                     -> dno
schedno_atom                -> dno_atom
tabno_atom                  -> dno_atom
DI                          -> dno_atom
DI C DI                     -> dno_atom
DI C DI                     -> dno_mol
bano dno                    -> dno_mol
schedno_atom t1_atom        -> dno_mol
schedno_mol t1_atom         -> dno_mol
schedno_atom t1_mol         -> dno_mol
schedno_mol t1_mol          -> dno_mol
dno_mol dno_atom            -> dno_mol
schedno                     -> bano
tabno                       -> bano
```

Abb. 3.3.2.9 Regeln für eine vc_DDC-Grammatik

Definition von DDC-Notation in der erweiterten Backus-Naur-Form (Extended Backus-Naur Form, kurz EBNF):

```
C          ::= „“
DI         ::= „0“ | „1“ | „2“ | „3“ | „4“ | „5“ | „6“ | „7“ | „8“ | „9“
DO         ::= „.“
bano       ::= schedno | tabno
clano      ::= schedno | tabno
dno        ::= dno_atom | dno_mol | dno_head dno_tail
dno_atom   ::= DI | DI C DI | schedno_atom | tabno_atom
dno_head   ::= sm | sd | ss | s
dno_mol    ::= DI C DI | bano dno | schedno_atom t1_atom | schedno_mol
               t1_atom | schedno_atom t1_mol | schedno_atom t1_atom |
               dno_mol dno_atom
dno_span   ::= dno_span1-dno_span2
dno_span1  ::= dno
dno_span2  ::= dno
dno_tail   ::= [ DI C DO C ]* DI ([C DI ])*
s          ::= ss C DO C dno_tail
sd         ::= sm C DI
sm         ::= DI
ss         ::= sd C DI
schedno    ::= dno | dno_span | s
tabno      ::= dno | dno_span
```

Abb. 3.3.2.10 DDC-Notation-Definition in EBNF

3.3.3 Algorithmus zur DDC-Notationsanalyse

Liu gibt in seiner Dissertation in Kapitel 4B. (Regeln zur Zerlegung) drei Arbeitsschritte (Kapitel 4B.1.[77]) an, die bei der Zerlegung einer synthetischen DDC-Notation beteiligt sind:

„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."

Abb. 3.3.3.1 Liu's Schritte zur Zerlegung einer synthetischen DDC-Notation

Liu veranschaulicht im Weiteren[78] für jede seiner 17 Regeln zur Zerlegung[79] die durchzuführenden Schritte zur Anwendung seiner Regel(n) mit je einem Beispiel (bis auf einen Fall) aus seiner Stichprobe. Hier wird ein Beispiel („in_liu_85") wiedergegeben, welches die Regeln A1 und Regel A2 verwendet:

„RULE A1: If AN1, concatenate the number after ‚the numbers following' and the target number minus the base number[80]; insert a decimal after the third digit, if the concatenated number has more than three digits; match the number against the Schedule."

Abb. 3.3.3.2 Liu's A1-Regel zur Zerlegung

„RULE A2: If AN2, match the target number minus the base number against the Schedule. If the number to be matched has two digits or fewer, add zeros to the end of the number until a number with 3 digits is reached."

Abb. 3.3.3.3 Liu's A2-Regel zur Zerlegung

"Example: Decompose 743.8979133 (How to draw clowns)"

Abb. 3.3.3.4 Liu's Beispielnotation „in_liu_85"

[77] [Liu1993], pp. 39-57 „IV. Findings, B.1. The procedure for Decomposing Numbers"
[78] Kapitel 4B.2. „Rules for Add Notes" und Kapitel 4B.3. „Rules for Standard Subdivisions
[79] hier in den Abbildungen Abb. 3.2.5 und Abb. 3.2.6 wiedergegeben
[80] „The target number refers to the complete number to be matched. The base number refers to the number immediately following the character string 'Add to [base number]' in a note. The target number minus the base number is the target number with the base number removed."

„1. Match the number 743.8979133 against the Schedule by dropping digits on the right; this retrieves the class 743.8 (Drawing other subjects). The class has Add Note AN1: 'Add to base number 743.8 the numbers following 704.94 in 704.943-704.999, e.g., ...'
2. Apply Rule A1 to match 704.94979133 against the Schedule.
3. Retrieve 704.949 (Other specific subjects) which has the Add Note AN2: 'Add to base number 704.949 notation 001-999, e.g., ...'
4. Apply Rule A2: match 791.33 (the target number 704.94979133 minus the base number 704.949) against the Schedule.
5. Retrieve the class 791.33 (Clowns).
6. Conclude that 743.8979133 decomposes to 743.8, [704.94]9, and 791.33."

Abb. 3.3.3.5 Liu's Zerlegungsschritte für die DDC-Notation „743.8979133"

Sein Programm DND (Dewey Number Decomposer) gibt hierzu aus[81] :

„743.8979133 has been decomposed as follows:
743.8: Drawing other subjects
(743: Drawing and drawings by subject)
704.949: Other specific subjects
(709.94: Subjects)
791.33: Clowns
(791.3: Circuses)
The title of this book is:
245 10#allow to draw clowns #cwritten and illustrated by Barbara Levy.
The subject headings for this book are:
#aClowns in art #xJuvenile literature.
#aHuman figure in art #xJuvenile literature.
#aDrawing #xTechnique #xJuvenile literature.
#aClowns in art.
#aDrawing #xTechnique.

Abb. 3.3.3.6 Liu's DND-Ausgabe für die „zerlegte" DDC-Notation „743.8979133

Eine – an Liu's Vorgehensweise eng angelehnte – 1. Version einer Analysekomponente („ul-ana.awk", Entwicklungszeitraum 18.03.04 – 31.08.04) im Rahmen von Colibri liefert:

```
743.8979133 <ddc_no_85_unabridged>
743.------- drawing and drawings by subject <ddc_no_atomar_from_schedule>
743.8------ drawing other subjects <ddc_no_abridged>
---.-979133 <ddc_no_tail>
---.-9----- other specific subjects <ddc_no_atomar_from_table>
---.--79133 clowns <ddc_no_atomar_from_table>

*** [main] analyse of ddc_no "743.8979133" is finished ***
```

Abb. 3.3.3.7 "ul-ana.awk"-Ausgabe für die analysierte DDC-Notation „743.8979133"

[81] [Liu1993] p. 111

Das awk-Programm „ul-ana.awk“ („analyze synthesized dewey decimal classification numbers“) umfasst zu diesem Zeitpunkt 1378 Zeilen („lines of code“), hinzu kommen 405 Zeilen des Vorverarbeitungsprogramm „ul-e22-pre“ („preparation of file E22.xml“), welches die DDC-Wissensbasis als Eingabedatei für „ul-ana.awk“ erzeugt. Von den 600 DDC-Notationen (Testdatei „in_liu“, vgl. Fußnote 74) werden 454 DDC-Notationen korrekt (d. h. in inhaltlicher Übereinstimmung mit Liu's DND-Ergebnissen) analysiert.

Dieses 1. VZG-Colibri-Analyse-Ergebnis über Liu's gesamten DDC-Notation-Testbestand wird in Hinblick auf Korrektheit (Vergleich mit Liu's Zerlegungsergebnissen), Vollständigkeit (Erweiterung auf alle DDC-Klassen) und Effizienz (Laufzeit für DDC-Notationsanalyse) einer Kritik unterworfen[82]. Es wird untersucht, woran es liegt, dass einige Zahlen mit „ul-ana.awk“ überhaupt nicht (korrekt) analysiert werden können. Zum einen liegt es daran, dass Liu's DDC-Notationen entweder einen Druckfehler enthalten oder sich (auf Grund von Auflagenveränderungen) nicht mehr im DDC-System[83] befinden. Zum anderen stellt sich heraus, dass die von Liu vorgenommene Vorgehensweise (1. und 2. Schritt in Abb. 3.3.3.1) einige Nachteile aufweist bzw. an ihre Grenzen stösst. Z. B. werden manche DDC-Notationen zu wenig tief, überhaupt nicht oder inkorrekt analysiert. Eine Ursache für diese Fehlerquelle liegt in den in den Tafeln enthaltenen synthetischen DDC-Notationen, die zu Zeiten von Liu's Untersuchungen ev. noch nicht im DDC-System enthalten waren. Es gibt weiterhin Fälle, wo das DND-Ergebnis nach Liu's nur rudimentär beschriebenem Verfahren nicht rekonstruierbar ist, was anhand der DDC-Notation „780.710749“ (in_liu_127) gezeigt werden soll:

Der in Abb. 3.3.3.1 gegebene 1. Schritt führt zu dem DDC-Klassenfund: „780.71“ („Music--education“). Dort gibt es keine Anweisung, laut 2. Schritt (Abb. 3.3.3.1) wird dann Hilfstafel 1 verwendet und ergibt: „T1--074“ („Museums, collections, exhibits“) und mit folgender Anweisung: „Add to base number T1--074 notation T2--1-T2--9...“ erhält man schliesslich: „T2--9“ („Other parts of world and extraterrestrial worlds“). Diese DDC-Notationsanalyse ist inkorrekt. Der Fehler liegt in der Ermittlung (match) von „780.71“: es ist eine synthetische DDC-Notation. Angenommen, diese DDC-Notation war im Jahr 1993 nicht in der Haupttafel vorhanden, dann muss damals die DDC-Notation „780.7“ („Education, research, performances, related topics“) gefunden worden sein. An die DDC-Notation „780.7“ kann kein Notationsteil hinzugefügt werden, weder mit einer Anweisung, noch mit der Möglichkeit, Hilfstafel 1 anzuwenden (da auf „780.7“ keine „0“ folgt). Um zum richtigen Ergebnis zu kommen, muss bis zur DDC-Notation „78“ („Music“) zurückgegangen werden, was Liu's Ausführungen bzw. seiner Skizze eines Algorithmus' nicht zu entnehmen ist. Da bei der DDC-Notation „78“ keine Anweisung steht, müssen die DDC-Notationsteile aus Hilfstafel 1 stammen: „T1--07101-T1--07109“ („Geographic treatment“) führt zum Erfolg. Dort steht die Anweisung: „Add to base number T1--0710 notation T2--1-T2--9“, was schliesslich zum letzten, korrekten DDC-Notationsteil „T2--749“ („New Jersey“) führt.

[82] im Sinne Poppers: vgl. z. B. „4. Zur Theorie des objektiven Geistes ...10. Der Wert von Problemen ... ‚Wie kann man lernen, ein wissenschaftliches Problem zu verstehen?' ... indem man versucht, es zu lösen, und dabei scheitert. ... selbst seine Lösung kritisieren ... Damit ist ein erster Schritt in Richtung auf die Lokalisierung der Schwierigkeit getan. Und das bedeutet nichts anderes, als daß ein erster Schritt zum Verständnis des Problems getan ist.“ in Karl R. Popper: „Objektive Erkenntnis: ein evolutionärer Entwurf.“ 1. Aufl. Hoffmann und Campe Verlag, Hamburg, 1993, S. 186 ff

[83] Liu verwendet die 20. Auflage (Jahr 1993), VZG-Colibri die 22. Auflage (Jahr 2003) des DDC-Systems. Aufgrund der unterschiedlichen Auflagen gibt es Unterschiede in den Klassenbenennungen, Systemverlegungen etc.

Es gibt auch Fälle, wo DND an seine (vordefinierten) Grenzen stößt, nämlich wenn Synthese-Anweisungen („add notes“) DDC-Klassen enthalten, die aus der Hauptklasse „700“ hinaus in andere DDC-Klassen verweisen, DND meldet in solchen Fällen[84]: „The DND database contains only the schedule 700“.

Wegen der dargelegten Gründe werden Systemarchitektur, Algorithmus, Datenstruktur, Programmcode und -namen geändert bzw. verbessert. Eine 2. Entwicklungsphase (15.10.04 – 22.11.04) und 3. Entwicklungsphase (seit 25.11.04) schliessen sich an (z. Zt. ca. 1100 Zeilen awk-/ksh[85]-Programmcode). Die DDC-Notationsanalyse wird nun von links nach rechts durchgeführt. Der derzeitige Algorithmus der DDC-Notations-Analysekomponte wird nun zunächst in Pseudocode[86], weiter unten durch ein Programmfragment dargestellt:

```
BEGIN (analyzing DDC numbers [dno's])
  1. while (there are lines in ddc_no_facts) do begin
       read and split line, store each item into a single array element
       1. dno 2. class 3. caption
     end
  2. while (there are lines in ddc_no_rules) do begin
       read and split line, store each item into a single array element
       1. dno 2. class-note
       if (class-note is „<nfa>“) store each in a single array element
       3. bano 4. notation 5. exception 6. exception notation
       if (class-note is „<ba.>“) store each in a single array element
       3. heading 4. class-note 5. add_to 6. bano 7. notation 8. preno
       if (class-note is „<na.>“) store each in a single array element
       3. bano 4. notation 5. preno
     end
  3. Print "Welcome"
  4. while (there are dno's to analyze) do begin
       for the dno_to_analyze print all dno-caption pairs taken from
       ddc_no_facts, i.e. first lines of a vc_daygram, i.e. call function
       "get_facts"
       while (there is more to analyze) do_begin
       for the dno_to_analyze print all dno-caption pairs taken from
       ddc_no_facts via applicable ddc_no_rules, i.e. last lines of
       a vc_daygram. i.e. call function "get_facts_via_rules"
       end
     end
  5. Print „done“
END
```

Abb. 3.3.3.8 DDC number analyzer ***vc_day*** in Pseudocode

[84] z. B. [Liu1993], p. 116

[85] [Herold1999b], S. 249-476. Helmut Herold: "Linux-Unix-Shells. Bourne-Shell, Korn-Shell, C-Shell, bash, tcsh.“ 3., akt. Aufl., Addison-Wesley, 1999. Kapitel 5: "Die Korn Shell." und [RosenblattRobbins2002] Bill Rosenblatt; Arnold Robbins: Learning the Korn Shell. 2nd ed. O'Reilly, Beijing, Cambridge u.a., 2002

[86] [AhoHopcroftUllman1983] Alfred Aho; John E. Hopcroft; Jeffrey D. Ullman: "Data Structures and Algorithms." Addison-Wesley, Reading u.a., 1983, p. 2: „In addition to using Pascal programs as algorithms, we shall often present algorithms using a *pseudo-language* that is a combination of the constructs of a programming language together with informal English statements."

```
function get_facts (for the dno_to_analyze)
  1. Determine the length of the dno_to_analyze
     [dno_to_analyze is considered as an abstract data type “list”,
      consisting of dno_head and dno_tail, dno_tail shrinks with
      growing dno_head]
  2. set length of dno_head to “0” (zero)
  3. while (there is a dno_tail) do begin
     3.1. increment length of dno_head by one
     3.2. determine dno_head
     3.3. determine dno_head’s caption
     3.4. print one line of daygram
     end
end
```

Abb. 3.3.3.9 ***get_facts*** in Pseudocode

```
function get_facts_via_rules (for the dno_to_analyze)
  1. Determine the length of the dno_to_analyze
  2. set length of dno_head to “0” (zero)
  3. while (there is a dno_tail) do_begin
     3.1 determine dno_head, dno_tail, dno_part
     3.2 if (there is a rule or table1 to apply)
         apply, if possible, rule or table 1, i.e. call function
         “get_and_apply_rule” (for dno_part)
     3.3 adjust dno_head, dno_tail, ...
    end
end
```

Abb. 3.3.3.10 ***get_facts_via_rules*** in Pseudocode

```
function get_and_apply_rule (get and apply the rule for the dno_head)
  if (there is a rule for the dno_head)
    if (it is a nfa-rule)
      get_and_apply_rule (recursive !)
      else
    if (it is a ba-rule)
      apply_ba
      else
    if (it is a na-rule)
      apply_na
  else (if there is a rule for a span of numbers)
    for (all dno_head in class-note)
      if (dno_head is located in a span of numbers and)
        if (it is a nfa-rule)
          get_and_apply_rule (recursive !)
          else
        if (it is a ba-rule)
          apply_ba
          else
        if (it is a na-rule)
          apply_na
  end
  apply_tb1 (if applicable)
end
```

Abb. 3.3.3.11 ***get_and_apply_rule*** in Pseudocode

Es folgt der Algorithmus zur Analyse der DDC-Notationen als Programmfragment in awk-Syntax:

```
BEGIN {
  while (getline ddc_no_facts_line < ddc_no_facts) {
    split(ddc_no_facts_line,arr,SUBSEP); # $1:dno $2:<hat|ri> $3:caption
    ...
  }
  while (getline ddc_no_rules_line < ddc_no_rules) {
    split(ddc_no_rules_line,arr,SUBSEP); # $1:dno $2:xtags
    ...
    if (substr(arr[2],1,4) == "<nfa") { #$3:fbanos $4:fnotations
                                        #$5:fexcepts $6:fspecs
      ...
    } else {
      ...
      if (substr(arr[2],1,3) == "<ba") { #$3:arehs $4:axtags $5:add_tos
                                         #$6:abanos $7:anotations
                                         #$8:aprenos
        ...
      } else {
        if (substr(arr[2],1,3) == "<na") { # $3:banos $4:notations
                                           # $5:prenos
          ...
        }
      }
    }
  }
  close(ddc_no_facts);
  close(ddc_no_rules);
  ...
  dno_atom_end = “”; # initialization of dno_atom_end
                     # dno_atom_end: not (or last) printed dno_atom
                     # used in get_dno_atoms
  printf("Welcome here ... this is work in progress ...\n");
  ...
# --------------------------- [main] beg----------------------------------#
  while (getline) {
    ...
    dno_to_ana = $1; # DDC number to analyze
# --------------------- analysing dno beg -------------------------------#
    ...
    get_facts(dno_to_ana);
    while (lt_save > 0 || lt > 0) { # there is more to analyze!
      lt_save = 0;
      get_facts_via_rules(dno_to_ana);
    }
    printf("\n*** d o n e ***\n");
    ...
# --------------------- analysing dno end -------------------------------#
  }
}
```

Abb. 3.3.3.12 awk-Hauptprogramm ***vc_day.awk*** (Fragment)

```
function get_facts(dno) { # FUGFA
  ld = length(dno); # ld: length of ddc number to analyze
  lh = 0;           # lh: length of dno_head
  lt = ld-lh;       # lt: length of dno_tail
  while (lt > 0) {  # as long as there is a dno_tail
    lh++;
    lt = ld-lh;
    dno_head = substr(dno,1,lh);
    ...
    if (lh != 4) {  # position "4" is dot
      get_caption(dno_head); # determine dno_head's caption
      ...
      pri_dno_cap(dno_head,caption,infotxt_pri,lh,lt) # print one line of
                                                      # daygram
      ...
    }
  }
}
```

Abb. 3.3.3.13 Benutzerdefinierte awk-Funktion „get_facts.awk“ (Fragment)

```
function get_facts_via_rules(dno_to_ana) { # FUGFARU
  ld = length(dno_to_ana);    # ld: length of ddc number to analyze
  lh = 0;                     # lh: length of dno_head
  lt = ld-lh;                 # lt: length of dno_tail
  ...
  while (lt > 0) {            # as long as there is a dno_tail
    if (dno_atom_end == “”) { # if 1st time or last printed dno_atom
                              # is empty
      lh++;
      lt = ld-lh;
      dno_head = substr(dno_to_ana,1,lh);
      dno_tail = substr(dno_to_ana,lh+1);
      dno_part = dno_head;
      ...
    } else {
      dno_part = dno_atom_end substr(dno_tail,1);
      ...
    }
    get_and_apply_rule(dno_part,r0_pri); # look whether there is an
    ...                                  # applicable rule
    if (dno_pri != “”) { # adjust dno_head, dno_tail, dno_part, ...
      dno_head = dno_head dno_pri;
      lh = length(dno_head);
      dno_pri = “”;
      dno_part = dno_atom_end;
    }
    ...
  } # eof-while
  ...
}
```

Abb. 3.3.3.14 Benutzerdefinierte awk-Funktion „get_facts_via_rules.awk“ (Fragment)

```
function get_and_apply_rule(dno_head_rule,r0_pri) { # FUGAPPRU
  if (xtags[dno_head_rule] != "") {  # there is a rule for the (single) dno
    if (xtags[dno_head_rule] == "<nfa>") {          # nfa-rule
      r0_pri = “nfa->”fnotations[dno_head_rule];
      get_and_apply_rule(fnotations[dno_head_rule],r0_pri); # recursive !
    } else
    if (substr(xtags[dno_head_rule],1,3) == "<ba") {# ba-rule
      apply_ba(dno_head_rule,r0_pri);
      ...
    } else
    if (substr(xtags[dno_head_rule],1,3) == "<na") {# na-rule
      apply_na(dno_head_rule,r0_pri);
    }
  } else {                          # is there a rule for span of numbers?
    for (dno in xtags) {
      get_span(dno);
      if (dno_span1 <= dno_head_rule && dno_head_rule <= dno_span2) {
        if (xtags[dno] == "<nfa>") {
          r0_pri = “nfa->”fnotations[dno];
          get_and_apply_rule(fnotations[dno],r0_pri); # recursive !
        } else
        if (substr(xtags[dno],1,3) == "<ba") {
          apply_ba(dno_head_rule,r0_pri);
          ...
        } else
        if (substr(xtags[dno],1,3) == "<na") {
          apply_na(dno,r0_pri);
        }
        ...
      }
    }
  ...
  apply_tb1(dno_to_ana);                      # is table 1 applicable?
  ...
  }
}
```

Abb. 3.3.3.15 Benutzerdefinierte awk-Funktion „get_and_apply_rule.awk“ (Fragment)

Zum erleichternden Verständnis werden in Abb. 3.3.3.16 die in Abbildungen Abb. 3.3.3.12 - Abb. 3.3.3.15 verwendeten Termini (ergänzend zu denen aus Abb. 3.3.27) zusammengefasst:

```
OCLC Legende (XML tags defined in E22.xml, ddc.dtd)
-------------------------------------------------------------------------
<ba[1345]>      beginning of add table
  <ba1>           part of schedule number
  <ba3>           part of table number
  <ba4>           all of table number
  <ba5>           identified by symbol
  <hat>           hierarchy at class
<na[1-6]>       add (footnote) note
  <na1>           part of schedule number
  <na2>           full schedule number
  <na3>           part of table number
  <na4>           add of table number
  <na5>           identified by symbol
  <na6>           referral to table
<nad>           generic instruction
<nfa>           add footnote
<nbu>           build note
<nfu>           modified standard subdivisions footnote
<nst>           modified standard subdivisions note
<reh>           heading in an add table
```

```
<ren>           notation in add table
<RI>            DCC Relative Index mapped intellectually by a DDC Editor
-------------------------------------------------------------------------
vc_day.awk Legende
-------------------------------------------------------------------------
SUBSEP                  SUBSEP = "\237"; # Ÿ
abanos                  a: add table, base number
add_tos                 a: add table, add_to number
anotations              a: add table, ... in notation
aprenos                 a: add table, preno (see below)
arehs                   a: add table, heading in add table
arr                     array
axtags                  a: add table, XML tags
banos                   base number(s)
daygram                 DDC number analysis diagram
ddc_no_facts            input file for "vc_day.awk"
ddc_no_facts_line       variable contains 1 line of ddc_no_facts
ddc_no_rules            input file for "vc_day.awk"
ddc_no_rules_line       variable contains 1 line of ddc_no_rules
dno                     DDC notation (number)
fbanos                  f: footnote, base number
fexcepts                f: footnote, ... except ...
fnotations              f: footnote, ... in notation
fspces                  f: footnote, special number (class ... in ...)
notations               notation
prenos                  ... the numbers following <preno> in ...
ri                      Relative Index Term
xtags                   XML tags
-------------------------------------------------------------------------
gawk_Legende [aus man page gawk]
-------------------------------------------------------------------------
SUBSEP                  The character used to separate  multiple  sub-
                        scripts  in array elements, by default "\034".
close(file [, how])     Close file, pipe or co-process...
getline var <file       Set var from next record of file.
printf fmt, expr-list   Format and print.
split(s, a [, r])       Splits the string s into the array a on the regular
                        expression r, and returns the number of fields. ...
substr(s,i[,n])         Returns the at most n-character substring of s
                        starting at i. If n is omitted, the rest of s is used
```

Abb. 3.3.3.16 Legende für OCLC- / vc-day.awk- / gawk-Termini (s. auch Abb. 3.3.2.7)

Wie in Abb. 3.3.2.3 wiedergegeben und oben erwähnt, wird die DDC-Wissensbasis ***vc_KB*** aus der DDC-Datei „E22.xml“ mit der Teilkomponente ***vc_ckb*** des DDC-Suchsystems ***vc_ds*** erstellt. Die aus den beiden Dateien „ddc_no_facts“ und „ddc_no_rules“ bestehende DDC-Wissensbasis ***vc_KB*** wird von der Teilkomponente ***vc_day*** zeilenweise vollständig eingelesen. Hauptaufgabe von ***vc_ckb*** (create knowledge base) ist die Extraktion der für die DDC-Notationsanalyse relevanten Daten aus „E22.xml“ (wohlgeformtes, gültiges XML-Dokument). Als Inhalt von XML-Elementen sind diese Daten in einen Text der natürlichen Sprache eingebettet. Die DDC-Fakten[87] sind aus dem XML-Dokument mit dem regulären Ausdruck: Re1: „<hat>|<RI>“ zu gewinnen, die aus den Anweisungen erzielbaren DDC-„Add“-Regeln[88] mit dem regulären Ausdruck Re2: „<na[1-6]>|<nad>|<nfa>|<ba[1345]>“.

[87] 48.067 („<hat>“) + 95.319 („<RI>“) = 143.386 DDC-Fakten

[88] 918 („<na1>“) + 79 („<na2>“) + 312 („<na3>“) + 492 („<na4>“) + 20 („<na5>“) + 57 („<na6>“) + 10 („<nad>“) + 5.381 („<nfa>“) + 9 („<ba1>“) + 3 („<ba3>“) + 12 („<ba4>“) + 75 („<ba5>“) = 7.368 DDC-Regeln

Die für die DDC-Notationsanalyse mindestens notwendigen 7.368 DDC-„Add“-Regeln können auch mit folgendem regulären Ausdruck Re3 beschrieben werden:

```
colibri/ul-test> 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 | wc
   7705  118938 1688088
```

Abb. 3.3.3.16 Regulärer Ausdruck zur Extraktion aller DDC-Anweisungen zur Konstruktion von DDC-Notationen aus der Datei „E22.xml“

Wie zu sehen ist, werden durch Re3 („`add to ... notation)`“) weitere 337 Zeilen ermittelt, die ggf. auch eine Rolle bei der Gewinnung von DDC-Regeln spielen. Es gibt weitere Anweisungen wie z.B. „<nbu>“, „<nfu>“ (vgl. Abb. 3.3.3.16). Nach Bestimmung der notwendigen Zeilen aus der XML-Datei „E22.xml“, extrahiert ***vc_ckb*** die für die DDC-Fakten und DDC-Regeln relevanten Werte aus dem Text der natürlichen Sprache (automatische Textanalyse) und gibt das Ergebnis in die beiden getrennten Dateien „ddc_no_facts“ bzw. „ddc_no_rules“ (:= ***vc_KB***) aus:

```
...
001.1Ÿ<ri>ŸIntellectual life (Fa1)
...
635.9528Ÿ<hat>ŸAlpine plants (Fa2)
...
```

Abb. 3.3.3.17 Zwei ***vc_KB***-Fakten aus „ddc_no_facts“

```
...
781.621-781.629Ÿ<ba4>ŸŸŸŸ781.62ŸT5—1-T5--9 (Ru1)
781.621-781.629+001-008Ÿ<ba4>ŸStandard subdivisionsŸ<nst>Ÿ00Ÿ780.1-780.9
                      Ÿ780 (Ru2)
...
808.819Ÿ<na4>Ÿ808.819ŸT3C--1-T3C--3 (Ru3)
...
```

Abb. 3.3.3.18 Drei ***vc_KB***-Regeln aus „ddc_no_rules“

Der ***vc_KB***-Inhalt kann folgendermassen (als Fakten und Regeln) gelesen werden:

Fa1: „Intellectual life“ ist ein zu der DDC-Klasse „001.1“ gehörender Fachgebietsbegriff aus dem DDC-Register („Relative Index Term“).

Fa2: „Alpine plants“ ist die Klassenbenennung der DDC-Klasse „635.9528“.

Ru1: Wenn es sich um eine innerhalb des DDC-Notationsbereiches „781.621-781.629“ liegende DDC-Notation handelt, dann ist es erlaubt, zu der Grundnotation „781.62“ Notationsteile aus dem Hilfstafel5-Notationsbereich „T5--1-T5--9“ hinzuzufügen.

Ru2: Wie Ru1, danach dürfen im Bereich „001-008“ liegende DDC-Notationsteile wie unter 780.1-780.9 (<nst>: modifizierte Hilfstafel 1-Notation) angegeben, hinzugefügt werden.

Ru3: Es ist erlaubt, zur Grundnotation „808.819“ Notationsteile aus dem Hilfstafel 3 – Notationsbereich „T3C--1-T3C--3 anzufügen.

4. Vergleich der Ergebnisse der Notationsanalysen (DND, *vc_day*)

Zur Veranschaulichung von durchgeführten Notationsanalysen greifen wir exemplarisch vier Beispiele aus Liu's Stichprobe (liu_347_to_analyze, liu_348_to_analyze, liu_529_to_analyze, liu_581_to_analyze) heraus, stellen die Zerlegungsergebnisse erst mit **DND** dann mit ***vc_day*** dar:

```
"781.62410763 has been decomposed as follows:
  781.621-781.629: Folk music of specific racial, ethnic, national groups
  41: French
  03-09: Specific continents, countries, localities
  763: Louisiana
 The title of this book is:
  245 10#aCajun music and Zydeco #cphotographs by Philip Gould: with
 The subject headings for this book are:
  #aMusicians, Cajun #zLouisiana #xPortraits
  #aZydeco musicians #zLouisiana #xPortraits"
```

Abb. 4.1 DND-Notationszerlegung für „liu_347_to_analyze" [Liu1993], S. 172

```
781.62410763 <liu_347_to_analyze>
7---------- Arts & recreation
78--------- Music
781-------- General principles and musical forms
781.6------- Traditions of music
781.62------ Folk music
781.6241---- French folk music
---.--4----- Modern Latin peoples <ba4:T5--4>
---.--41---- French <ba4:T5--41>
---.--410763 Cajuns <ba4:T5--410763>
---.----0--- <ba4:T5--41:na4:0>
---.-----7-- North America <nad:T2--7>
---.-----76- South central United States
             Gulf Coast states <nad:T2--76>
---.-----763 Louisiana <nad:T2--763>
```

Abb. 4.2 vc_daygram für „liu_347_to_analyze" mit Informationsausgabe zur Synthese

```
"781.626872079494 has been decomposed as follows:
  781.621-781.629: Folk music of specific racial, ethnic, national groups
  687-688: National groups
  72: Middle America####Mexico
  03-09: Specific continents, countries, localities
  79494: Los Angeles
 The title of this book is:
  245 10 #aBarrio rhythm: #bMexican-American music in Los Angeles
 The subject headings for this book are:
 #aMexican-Americans #zCalifornia #zLos Angeles #xMusic #xHistory
 #aPopular music #zMexico #xHistory and criticism
 #aPopular music #zUnited States #zLos Angeles"
```

Abb. 4.3 DND-Notationszerlegung für „liu_348_to_analyze" [Liu1993], S. 173

```
781.626872079494 <liu_348_to_analyze>
7--------------- Arts & recreation
78-------------- Music
781------------- General principles and musical forms
781.6----------- Traditions of music
781.62---------- Folk music
---.--6--------- Spanish and Portuguese <ba4:T5--6>
---.--68-------- Spanish Americans <ba4:T5--68>
---.--687------- Regional and national groups <ba4:T5--687>
---.--6872------ Mexicans <ba4:T5--6872>
---.------0----- <ba4:T5--6872:na4:0>
---.-------7---- North America <nad:T2--7>
---.-------79--- Great Basin and Pacific Slope of United States
                 Pacific Coast states <nad:T2--79>
---.-------794-- California <ba4:na4:T2--794>
---.-------7949- Southern California <nad:T2--7949>
---.-------79494 Los Angeles <nad:T2--79494>
```

Abb. 4.4 vc_daygram für „liu_348_to_analyze“ mit Informationsausgabe zur Synthese

```
“781.62969729107471 has been decomposed as follows:
  781.621-781.629: Folk music of specific racial, ethnic, national groups
   (781.62: Folk music)
  969: Other national groups of largely African descent
  7291: Cuba
  03-09: Specific continents, countries, localities
  7471: New York####Borough of Manhattan (Manhattan Island,New York County)
 The title of this book is:
  245 10 #aSalsiology: #bAfro-Cuban music and the evolution of salsa in
 The subject headings for this book are:
  #aSalsa #xHistory and criticism.
  #aPopular music #zCaribbean Area #xHistory and criticism.
  #aPopular music #zNew York (N.Y.) #xHistory and criticism.
```

Abb. 4.5 DND-Notationszerlegung für „liu_529_to_analyze” [Liu1993], S. 215

```
781.62969729107471 <liu_529_to_analyze>
7---------------- Arts & recreation
78--------------- Music
781-------------- General principles and musical forms
781.6------------ Traditions of music
781.62----------- Folk music
---.--9---------- Other ethnic and national groups <ba4:T5--9>
---.--96--------- Africans and people of African descent <ba4:T5--96>
---.--969-------- Other regional and national groups of largely African
                  descent <ba4:T5--969>
---.-----7------- North America <na4:T2--7>
---.-----72------ Middle America    Mexico <na4:T2--72>
---.-----729----- West Indies (Antilles) and Bermuda <na4:T2--729>
---.-----7291---- Cuba <na4:T2--7291>
---.---------0--- <ba4:na4:0>
---.----------7-- North America <na4:T2--7>
---.----------74- Northeastern United States (New England and
                  Middle Atlantic states) <na4:T2--74>
---.----------747 New York <na4:T2--747>
---.----------7471 New York    Borough of Manhattan (Manhattan Island, New
                  York County) <na4:T2--7471>
```

Abb. 4.6 vc_daygram für „liu_529_to_analyze“ mit Informationsausgabe zur Synthese

Die in den Abbildungen Abb. 4.2, Abb. 4.4. und Abb. 4.6 angegebenen Ausgaben als ***vc_daygram*** können auch als ***vc_dayset*** (Beispielaufruf vc_ds -k vc_KB -i in_liu_529 -o vc_dayset) ausgegeben werden. Diese Menge kann mit folgender Anfrage ermittelt werden: „Welche Klassenbenennungen *cap* sind alle (hier symbolisiert durch „ ⊕ “) in der DDC-Klasse 781.62969729107471 enthalten?“[89]

$$I((\lambda x)\, dno\_cap(\oplus 781.62969729107471, x)) :=$$

{Africans_and_people_of_African_descent, Arts_&_recreation, Cuba, Folk_music,
General_principles_and_musical_forms, Middle_America____Mexico, Music, New_York,
New_York____Borough_of_Manhattan_(Manhattan_Island,_New_York_County),
North_America,Northeastern_United_States_(New_England_and_Middle_Atlantic_states),
Other_ethnic_and_national_groups,Other_regional_and_national_groups_of_
largely_African_descent,Traditions_of_music, West_Indies_(Antilles)_and_Bermuda}

Danach ist die Bedeutung der DDC-Notation 781.62969729107471 die Menge (***vc_dayset***) :

{Africans_and_people_of_African_descent, ..., West_Indies_(Antilles)_and_Bermuda} oder
{7,78,781,781.6,781.62,T5--9,T5--96,T5--969,T2--7,T2--72,T2--729,T2--7291,T2--7,T2--74,
T2--747,T2--7471}

Im **vc_daygram** wird - falls gewünscht - folgende Information in spitzen Klammern "<...>" zusätzlich ausgegeben (mit OPTION "-p explain", vgl. Abb. 3.3.2.5):

<[r1[->tabno|schedno]:[r2[->tabno|schedno]:r3:...]tabno>

mit ri (i=1, 2, ...) für eine angewendete ba-, na-, nfa-, ... -Regel.

Erläuterung: eine nfa-Regel verweist auf eine bei einer anderen DDC-Notation stehende Anweisung mit der Phrase: "Add as instructed under". Dieser Verweis wird im ***vc_daygram*** durch "->" dargestellt (in „get_and_apply_rule.awk“ durch Rekursion realisiert).

Zusätzlich zu den im Anhang seiner Arbeit abgedruckten DND-Zerlegungsausgaben (wie in Abb. 4.1, Abb. 4.3 und Abb. 4.5) gibt Liu in seiner Arbeit zwei Benutzer-Bildschirmausgaben („Sample DND Decomposition Screen“ und „Sample Explanation Screen“) an, die hier wiedergegeben werden:

```
System     Browse     Search     Decompose     Report     Adjust     Quit

--------------------------- Decompose DDC Number ------------------------

DDC Number:   796.323640979494

 796.323640979494 has been decomposed as follows:
 796.32364: Professional and semiprofessional basketball
 093-099: Treatment by specific continents, countries, localities; extra
 79494: Los Angeles
-------------------------------------------------------------------------
                         Why?              Exit
```

Abb. 4.7 DND-Bildschirmausgabe für „liu_581_to_analyze” [Liu1993], S. 24

[89] Anfragesprache analog IQL-DIT (Reiner 1991, S. 121ff, s. a. *Exkurs1*, *Exkurs2*) aufbaubar

Zusätzlich zur Bildschirmausgabe wie in Abb. 4.7 kann DND dem / der Benutzer/in auf die Frage „warum“ (Menuauswahl „Why“ in letzter Zeile) die einzelnen Zerlegungsschritte zur Erklärung ausgeben:

```
--------------------------- Decompose DDC Number ------------------------

DDC Number:   796.323640979494

----------------------------- Explanation Window ------------------------

Step 1: Schedule entry 796.32364 retrieved. Digit(s) 0979494 left
Step 2: 0979494 was searched against Table 1 because there is no
        Add note provided for entry 796.32364, and 796.32364 has one
        of the forms that belong to Standard Subdivision type 3
Step 3: Table 1 entry 093-099 retrieved. Digit(s) 9494 left
Step 4: Add Note found under 093-099:
        Add to base number 09 notation 2_3-2_9 from Table 2, e.g.,
        the subject in United States 1_0973, in Brazil, 1_0981, in
        North America 1_097; then add further as follows
Step 5: 79494 retrieved

---------------------- Press <Esc> key to exit --------------------------
```

Abb. 4.8 DND-Erklärungs-Bildschirm für „liu_581_to_analyze” [Liu1993], S. 25

Damit hat Liu im „Dewey Number Decomposer „ (DND) alle Komponenten eines Expertensystems[90] realisiert: Erwerbungs- und Erklärungskomponente, Wissensbasis und Inferenzmaschine. Auch die Teilkomponente ***vc_day*** des DDC-Suchsystems ***vc_ds*** kann als Expertensystem aufgefasst werden: Die Erwerbungskomponente wird durch ***vc_ckb*** realisiert, die die Wissensbasis ***vc_KB*** erzeugt. Die Inferenzmaschine ***vc_day*** gibt im ***vc_daygram*** implizit alle Information aus, um sich die Analyse-Schritte anzeigen (erklären) zu lassen (rudimentäre Erklärungskomponente):

```
796.323640979494 <liu_581_to_analyze>
7--------------- Arts & recreation
79-------------- Sports, games & entertainment
796------------- Athletic and outdoor sports and games
796.3----------- Ball games
796.32---------- Inflated ball thrown or hit by hand
796.323--------- Basketball
796.3236-------- Specific types of basketball
796.32364------- Professional and semiprofessional basketball
---.-----09----- Historical, geographic, persons treatment <T1--09>
---.-------7---- North America <ba4:T2--7>
---.-------79--- Great Basin and Pacific Slope of United States    Pacific
                 Coast states <ba4:T2--79>
---.-------794-- California <ba4:T2--794>
---.-------7949- Southern California <ba4:T2--7949>
---.-------79494 Los Angeles <ba4:T2--79494>
```

Abb. 4.9 vc_daygram für „liu_581_to_analyze“ mit Informationsausgabe zur Synthese

[90] vgl. z. B. [Forsyth1984] “Expert Systems: Principles and Case Studies.” (ed. Richard Forsyth), Chapman and Hall, London, 1984, p. 11

Ein ***vc_daygram*** (mit in „<...>“ enthaltener Informationsausgabe) enthält:

- die zu analysierende DDC-Notation (molekulare DDC-Notation),
- den Namen der zu analysierenden DDC-Notation
  (in 1. Zeile mit „<...>“ gekennzeichnet, z. B. „<liu_581_to_analyze>“),
- den Dewey-Punkt an 4. Stelle der DDC-Notation,
- die Reihenfolge und die Stellung der Ziffern der DDC-Notation,
- die relevanten DDC-Notationsteile pro Analyseschritt (atomare DDC-Notation),
- die Kennzeichnung der im Analyseschritt irrelevanten Stellen mit „-“,
- die Bedeutung der DDC-Notationsteile (Klassenbenennung) pro Analyseschritt,
- die angewendeten Regeln mit ermittelten DDC-Notationen aus den Hilfstafeln

Somit ist in einem ***vc_daygram*** Analyse- und Syntheseinformation enthalten. Eine „Erklärung“ zur Synthese der DDC-Notation „796.323640979494“ kann so gelesen werden (es könnte z. B. ein Text dieser Art ausgegeben werden):

Der Anfang der molekularen DDC-Notation „796.323640979494“ enthält einschliesslich bis zur Ziffernfolge „796.32364“ acht atomare DDC-Notationen mit Klassenbenennungen, die alle in der Haupttafel enthalten sind. Der verbleibende Notationsteil „0979494“ setzt sich folgendermassen zusammen: „09“ hat die Klassenbenennung „Historical, geographic, persons treatment“ und entstammt der Hilfstafel 1. Die atomare DDC-Notation „T2--7“ („North America“) entsteht durch Anwenden der bei der DDC-Notation „T1--07“ in einer Anhängetafel (ba4-Regel) stehenden Anweisung. Diese besagt, dass an die DDC-Notation „T1--09“ die „7“, die „79“ („Great Basin ...“), die „794“ („California“) etc. aus Tabelle 2 hinzugefügt werden kann (Originaltext in Anweisung: „Add to base number T1--09 notation T2--3-T2--9 from Table 2, ...“.).

Den Abschluss dieses Kapitels bildet die in Kapitel 3.3.2 (Fußnote 63) angekündigte DDC_Notationsanalyse:

```
635.9528095496 <ul_18_to_analyze>
6------------- Technology
63------------ Agriculture
635----------- Garden crops (Horticulture)    Vegetables
635.9--------- Flowers and ornamental plants
635.95-------- Groupings by environmental factors
635.952------- Groupings by climatic factors
635.9528------ Alpine plants
--------09---- Historical, geographic, persons treatment <T1--09>
----------5--- Asia    Orient    Far East <ba4:T2--5>
----------54-- South Asia    India <ba4:T2--54>
----------549- Other jurisdictions <ba4:T2--549>
----------5496 Nepal <ba4:T2--5496>
```

Abb. 4.10 vc_daygram für „ul_18_to_analyze“ mit Informationsausgabe zur Synthese

5. Zusammenfassung und Ausblick

1993 hat Liu in einem statistischen Test gezeigt, daß sich synthetische DDC-Notationen für die Klasse 700 („The arts Fine and decorative arts“) automatisch korrekt zerlegen lassen. Nach unserer Kenntnis steht das von ihm entwickelte Computersystem DND (Dewey Number Decomposer) nicht zur Verfügung, und der Programmcode samt Dokumentation ist nicht veröffentlicht worden. Im vorliegenden Bericht wird die Analysekomponente vc_day (des DDC-Suchsystems vc_ds) zur automatischen Notationsanalyse vorgestellt, mit dem sich dieTestergebnisse von Liu reproduzieren lassen. Es wird aufgezeigt, wie sich die Ergebnisse der Notationszerlegung zur Suche und zur Klassifikation verwenden lassen.

Die Unterschiede in den Zerlegungsergebnissen des Programmsystems DND bzw. der Analysekomponente vc_day beruhen zum einen auf den verwendeten DDC-Auflagen (20. bzw. 22. Auflage) des DDC-Systems. Zum anderen besteht ein wesentlicher Unterschied darin, dass - im Unterschied zum DND-System - in vc_day eine feinstmögliche Zerlegung einer molekularen DDC-Notation vorgenommen wird, nämlich in kleinste semantisch sinnvolle Zerlegungseinheiten (atomare DDC-Notationen).

Sowohl DND als auch vc_day können als DDC-Expertensystem aufgefasst werden, welche DDC-Experten bei der DDC-Notationsanalyse und -synthese unterstützen können. Als nächster Schritt soll die DDC-Notationsanalyse auf alle möglichen DDC-Notationen ausgedehnt werden. Die DDC-Notationen werden durch die (aufgezählten) Fakten und Regeln des DDC-Systems erzeugt, sie definieren implizit eine „Maschine zur DDC-Nummernerzeugung“[91]. Die DDC-Notationssynthese scheint eindeutig[92] und endlich[93] zu sein. Hieraus kann auf den Grammatiktyp der Sprache, die aus der Menge aller synthetischen DDC-Notationen besteht, geschlossen werden[94].

Im dritten Kapitel haben wir eine Grammatik zur Erzeugung von DDC-Notationen angegeben. Wir vermuten, dass die Grammatik regulär ist. Die Kenntnis des Grammatiktyps ist von praktischer Bedeutung für die Abschätzung der Laufzeit von Algorithmen.

[91] [ChanMitchell2003], p. 9 “The Dewey Decimal Classification is basically a number-building machine that provides both intellectual order and physical location in one step. The engine of the DDC is notation.”

[92] Es gibt allerdings Alternativen wie Option, Option A, Option B, Option C zur Notationsbildung (Optional number, vgl. Abb. 2.2)

[93] [ChanMitchell2003], p. 152 "Although the system does not allow for unlimited synthesis, many aspects of a subject can be expressed through number building."

[94] siehe z. B. http://wwwwbs.cs.tu-berlin.de/~ul/freiall.pdf, S. 9