

Librarian Workshop 8 March 2007

The 31st Annual Conference of the German Classification Society on Data Analysis, Machine Learning, and Applications, March 7-9, 2007, Freiburg i. Br., Germany

ENHANCEMENT OF UDC DATA FOR USE AND SHARING IN A NETWORKED ENVIRONMENT

Aida Slavic

Maria Ines Cordeiro

Gerhard Riesthuis



MAIN POINTS

- UDC facts update
- Logic behind the synthetic structure
- UDC number building and authority control
- Improvement of data
- Data development roadmap

UNIVERSAL DECIMAL CLASSIFICATION (UDC)

- classification system created to support
 - detailed document indexing in bibliographies
 - broad collocation of subject
- vocabulary – current database “UDC MRF - Master Reference File” contains 67.600 classes
- covers the whole universe of knowledge
- hierarchical, analytico-synthetic, faceted structure

UDC HIERARCHY

0 Generalities

1 Philosophy. Psychology

2 Religion

3 Social Sciences

4

5 Natural sciences

6 Applied sciences

7 Arts. Sports

8 Languages. Literature

9 Geography. History

0 Generalities. Organization. Information

00 Prolegomena. Fundamentals of knowledge and culture

005 Management

005.9 Fields of management

005.95 Personnel management

005.952 Workforce market

005.952.2 Staff ratios

5 Mathematics and natural sciences

53 Physics

539 Physical nature of matter

539.12 Elementary and simple particles

539.120 Theoretical problems of elementary particles physics

539.120.2 Symmetries of quantum physics

539.120.22 Conservation laws

539.120.222 Translations. Rotations

UDC FACTS

- used in at least 124 countries in the world
 - in 41 countries it is the main system
- printed editions exist in 39 languages
- available online in English, Czech, Swedish
- published on CD-ROM in Czech, Russian, Spanish

UDC FACTS

- the owner of the UDC is a not-for-profit association of publishers – UDC Consortium (<http://www.udcc.org>)
- classification is regularly maintained as the UDC Master Reference File (UDC MRF) database
- publications of the UDC are not centrally controlled
 - any stakeholder can purchase the licence to publish/translate UDC MRF or its parts

WHAT IS SPECIFIC TO UDC

- UDC is distributed to users as a database export file (UDC Master Reference File)
- Users can purchase a part or the whole of the vocabulary and implement it/map it to any other indexing language
- Users can re-use UDC MRF database and build their own local tool for classification

UDC MAINTENANCE AND DEVELOPMENT

Since 1993:

- intensive revision of the system
 - 14 editions in 14 years
- revision is based on the principle of facet analysis
- further improvement of the existing analytico-synthetic structure

FACET ANALYSIS

using an example of objects

BOXES

FACET 1
(criterion COLOUR)

red
yellow
blue

FACET 2
(criterion MATERIAL)

cardboard
wood
metal

FACET 3
(criterion SIZE)

small
medium
large

LINEAR PRESENTATION

NOTATION

- 1 BOXES
 - 11 **Cardboard Box**
 - 111 Red cardboard box
 - 112 Blue cardboard box
 - 113 **Yellow** cardboard box
 - 12 **Wooden Box**
 - 121 Red wooden box
 - 122 Blue wooden box
 - 123 **Yellow** wooden box
 - 13 **Metal Box**
 - 131 Red metal box
 - 132 Blue metal box
 - 133 **Yellow** metal box

enumeration

ENUMERATION OF COMPOUND CONCEPTS

1 BOXES

11 Cardboard Box

111 Red cardboard box

112 Blue cardboard box

113 Yellow cardboard box

12 Wooden Box

121 Red wooden box

122 Blue wooden box

123 Yellow wooden box

13 Metal Box

131 Red metal box

132 Blue metal box

133 Yellow metal box

2 SUITCASES

...

simple notation
expresses a
combination of
concepts/attributes

even with a small
number of objects
and properties the
size of the scheme
becomes an issue

CROSS CLASSIFICATION

1	BOXES	
11	Cardboard	
111		red
112		blue
113		yellow
114		large
115		medium
116		small
117		flat-pack
118		for food
12	Wooden	
121		red
122		blue
123		yellow
124		handmade
125		miniature
126		gilded
13	Metal	
131		red
132		blue
133		non-corrosive
134		golden
135		silver
126		with lockers

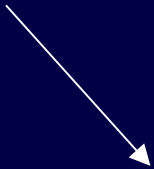
schemes makes a selection of useful combinations

hierarchy gets compressed to make notation shorter

into classes according to the mixture of criteria -> structural fault known as cross classification

HOW THIS WOULD LOOK IN UDC?

NOTATION
WITH FACET
INDICATORS



1 PRODUCTS

1-1 ACCORDING TO MATERIAL

- 1-11 Cardboard
- 1-12 Wooden
- 1-13 Metal

1.01 ACCORDING TO COLOUR

- 1.02 Red
- 1.03 Blue
- 1.04 Yellow

1'1 ACCORDING TO SIZE

- 1'12 small
- 1'13 medium
- 1'14 large

11 BOXES

12 VASES

13 SUITCASES

14 HANDBAGS

FACETED APPROACH
IN STRUCTURING
SCHEDULES

WHY IS THIS IMPORTANT?

no limitation to the combination of concepts

11-12.03'14 Boxes-wooden-blue-large

13-13.03'14 Suitcases-metal-blue-small

1 PRODUCTS

1-1 [FACET: MATERIAL]

1-11 Cardboard

1-12 Wooden

1-13 Metal

1.01 [FACET: COLOUR]

1.02 Red

1.03 Blue

1.04 Yellow

1'1 [FACET: SIZE]

1'12 small

1'13 medium

1'14 large

11 Boxes

12 Vases

13 Suitcases

14 Handbags

RESULTING IN ...

1 PRODUCTS

1-1 [FACET: MATERIAL]

1-11 Cardboard

1-12 Wooden

1-13 Metal

1.01 [FACET: COLOUR]

1.02 Red

1.03 Blue

1.04 Yellow

1'1 [FACET: SIZE]

1'12 small

1'13 medium

1'14 large

11 Boxes

12 Vases

13 Suitcases

14 Handbags

shorter schedules

more powerful indexing through
unlimited synthesis

postcoordinate searching with the
help of facet indicators

simple concepts are assigned simple
notation + symbol

easier linking of natural language
terms to notation

FACET ANALYSIS

an example of discipline/subject

LINGUISTICS

FACET 1
'SUBJECT'

English
German
French

FACET 2
'PARTS'

grammar
syntax
morphology

FACET 3
'PERIOD'

old
middle
modern

FUNDAMENTAL FACET CATEGORIES

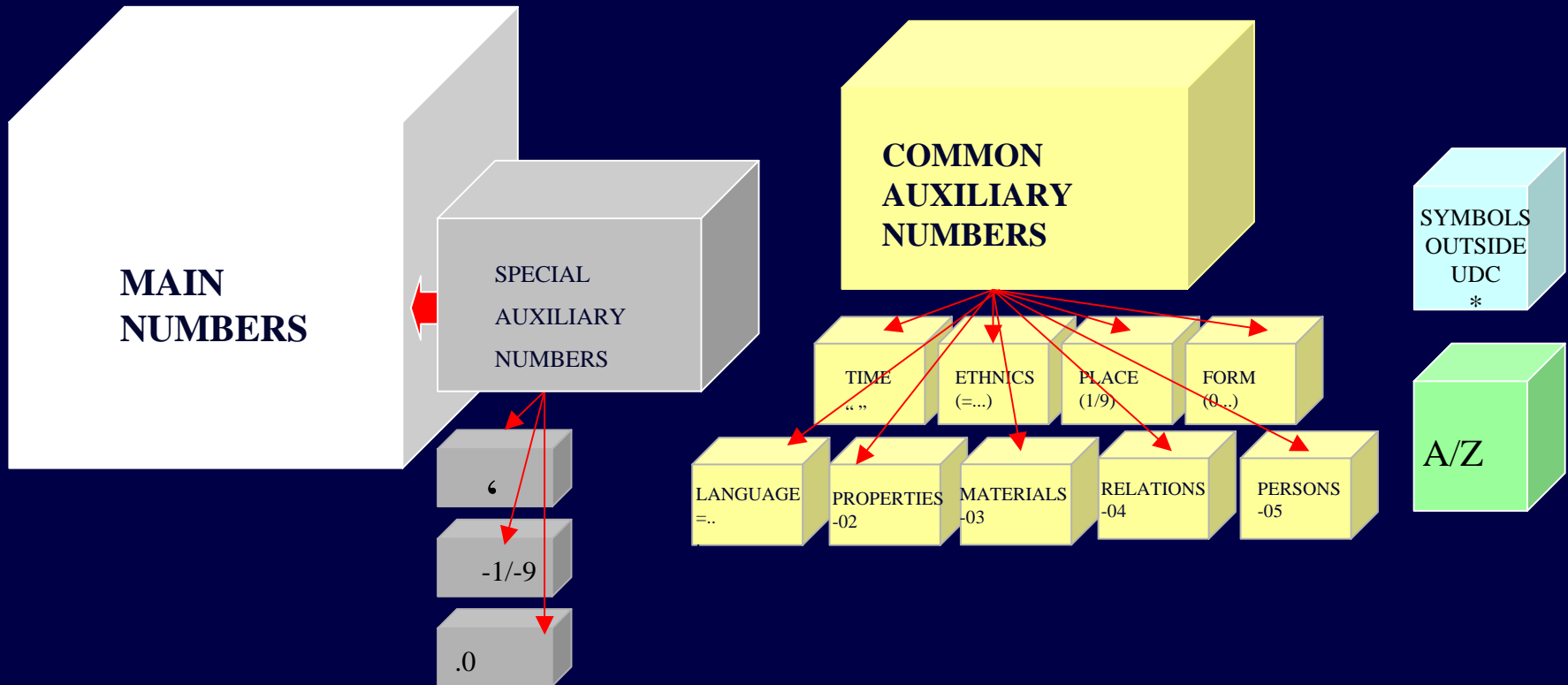
FUNDAMENTAL FACET CATEGORIES

- thing
 - kinds
 - parts
 - materials
 - properties
 - processes
 - operations
 - products
 - space
 - time
- e.g.
- plants
 - cereals
 - stem
 - cellulose
 - toxic, edible
 - growth
 - harvest
 - food, textile
 - desert, marshes
 - Middle Ages

FACETED AND SYNTHETIC FEATURES OF UDC

- allow for limited number of concepts to be used in indexing unlimited number of complex subjects
- allows more detailed indexing
- allows coordination in indexing and post-coordination in retrieval
- allows independent manipulation and use of mutually exclusive facets of vocabulary as independent access points
- allows non-linear presentation i.e. faceted interface

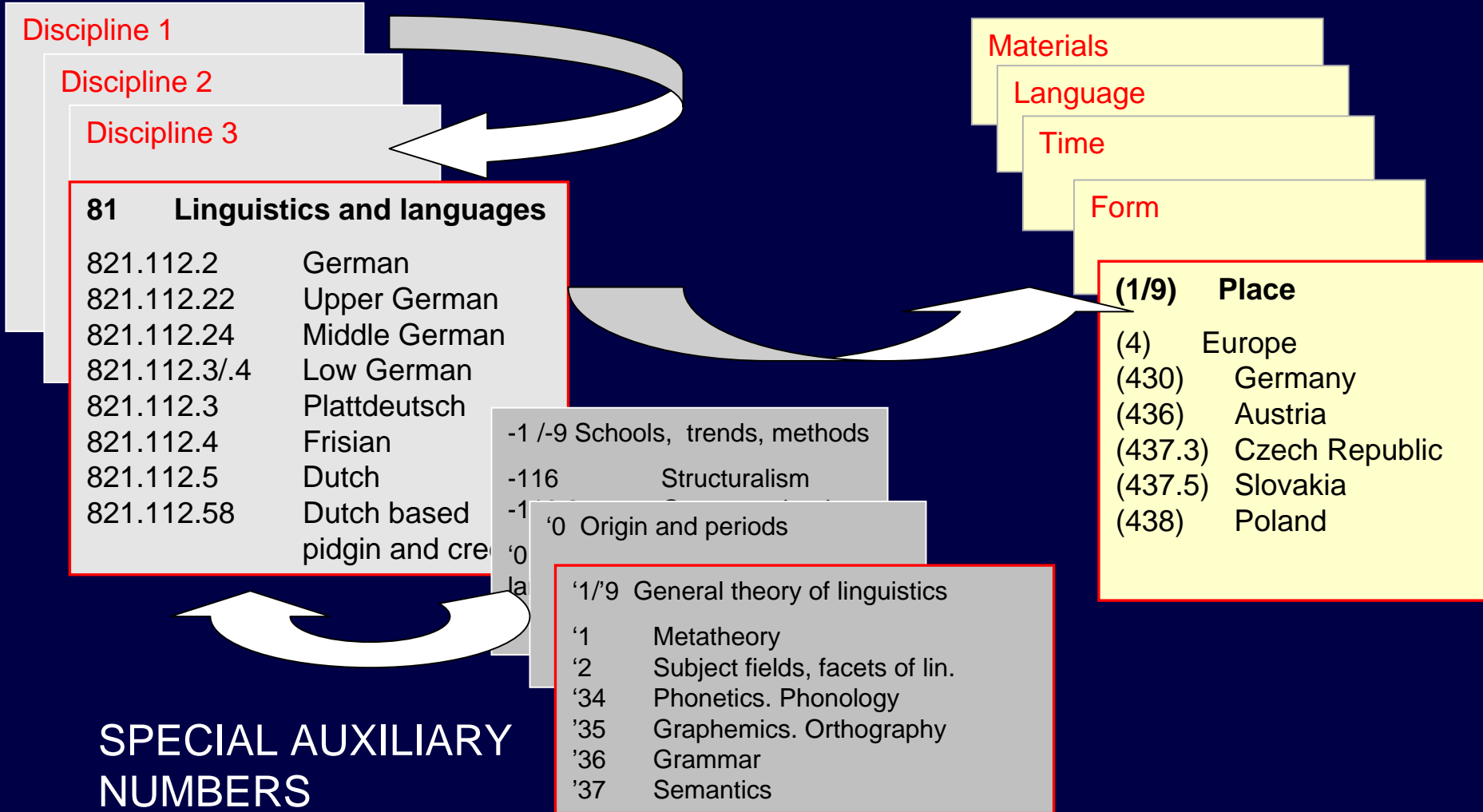
UDC – FULL PICTURE



SYNTHESIS IN UDC

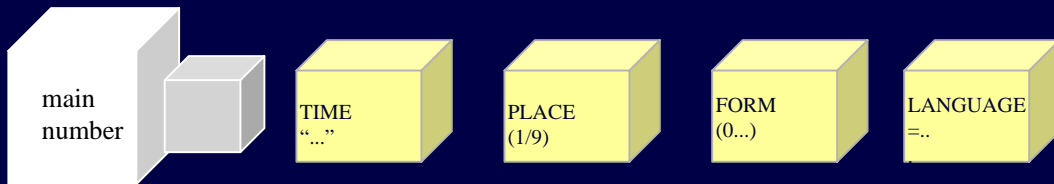
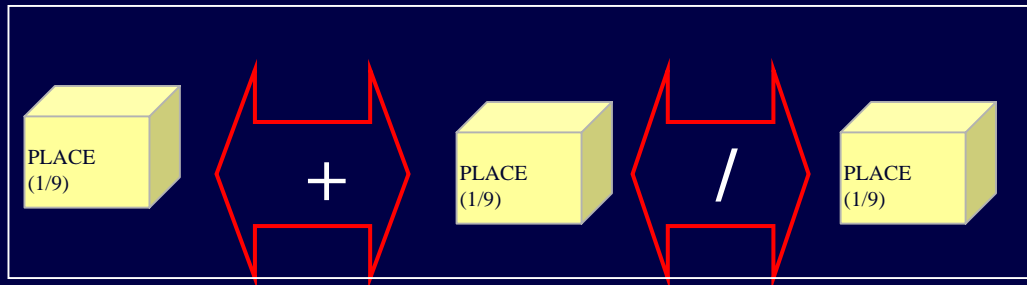
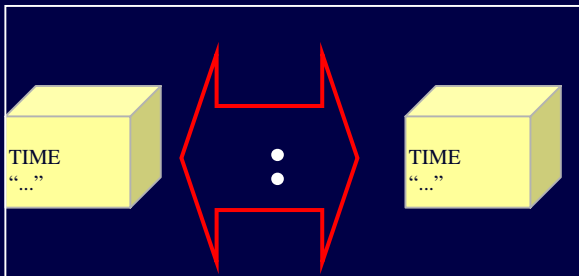
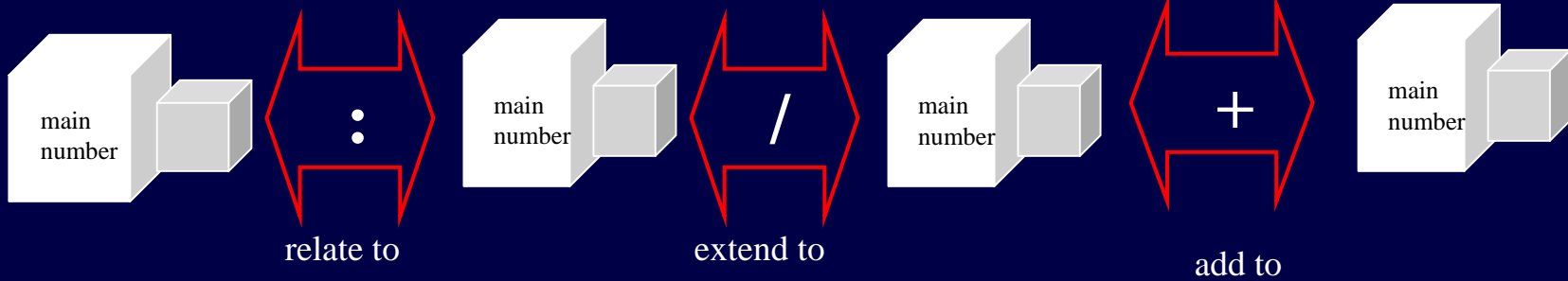
MAIN TABLES

COMMON AUXILIARY NUMBERS



SPECIAL AUXILIARY NUMBERS

SYNTAX



SYNTHESIS ACROSS DISCIPLINES

37 : 004

computer science

relationship

education

UNLIMITED SUBJECTS

338.48-6 : 502.172

eco- tourism
according to the motive

relationship

protection of the environment

SPEAKING UDC...

338.48(430-22) : 502.172

protection of the environment

relationship

?

tourism

Germany - rural

BEING VERY SPECIFIC ...

338.48(44-22)-042.3:502.172

**influence of
protection of the environment
on tourism
in Rural Germany**

FLEXIBILITY OF ORDER

94 (430) “18”

History Germany 19th century

94 “18” (430)

History 19th century Germany

FLEXIBILITY OF CHOICE

PLACE ONLY	(430.127.7)	Freiburg
PLACE/TIME	(430.127.7)"17"	Freiburg - 18 century
PLACE/TIME/Form	(430.127.7)"17"(089.7)	Freiburg – 18 th c. postcards
Form	(089.7) (430.127.3) (089.7) (430.127.7)	Postcards - Stuttgart Postcards – Freiburg

IMPLEMENTATION CHOICE

1. using UDC with simple notation
2. using UDC with synthesizeded, structured notation

SIMPLE NOTATION ONLY

hierarchy as presented in the main schedules/main facets

- 328** Parliaments. Representation of the people. Governments
- 328.1** Parliament and government
- 328.17** Resignation
- 328.172** Resignation of cabinet members
- 328.3** Function of parliament

COMPOSED (STRUCTURED) NUMBERS

328 Parliaments. Representation of the people. Governments

328"17" Parliaments in 18th century

328"18" Parliaments in 19th century

328"18"(4) Parliaments in 19th century in Europe

328.1 Parliament and government

328.1"18"(410) Parliament and government in 19th century in England

328.1"18"(430) Parliament and government in 19th century in Germany

328.1"18"(45) Parliament and government in 19th century in Italy

328.1"18"(5) Parliament and government in 19th century in Asia

328.1(410)(091) History of parliament and government in England

328.1(44)(091)=40 History of Parliament and government in France. In French language

328.17 Resignation

328.172 Resignation of cabinet members

328.3 Function of parliament

CHOICE WITH COMPOSED UDC NUMBERS

Building numbers in indexing but treating them as simple subjects in retrieval

328.1"18"(430)

German 19th century government

Preserving complex structure in indexing and in retrieval

328.1

"18"

(430)

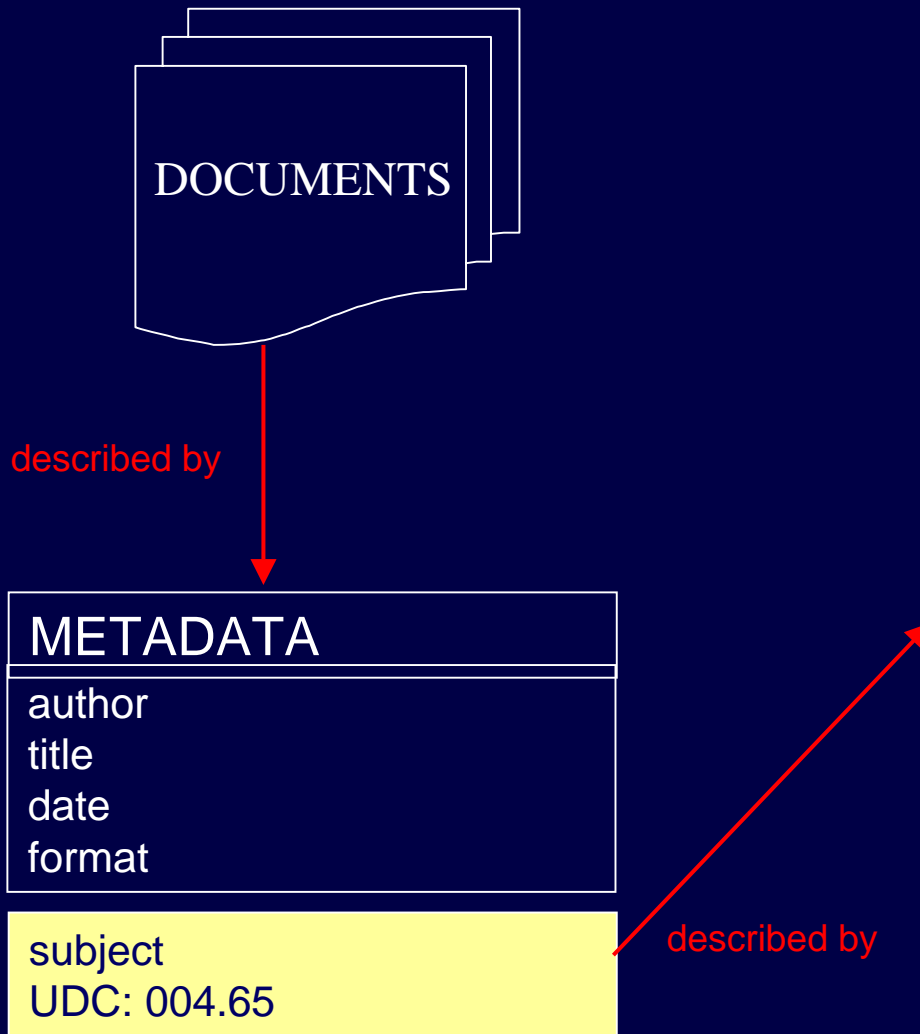
Government

19th century

Germany



CLASSIFICATION AUTHORITY CONTROL



UDC AUTHORITY FILE

heading: 004.64
description: Database management systems
was before: n.a.
Broader class: 004.6
Related classes:
Search terms: Database
DBMS
Database operating system
DB
display: 004.64

MAPPING

local classification:
LCSH
Dewey

CLASSIFICATION AUTHORITY FILE

used to support structured UDC headings, it allows

- separation of data content from its presentation
- multiple access points to classification
- semantic relationships between classes to be exploited for search expansion
- mapping of classification to other indexing systems (other classifications, thesauri, subject heading systems)
- management of global changes and reclassification

MAIN ASSUMPTIONS IN SUPPORTING UDC USERS

- that subject data already is and will increasingly be managed through authority files
- may want combine/expand UDC data with other special classification systems/thesauri
- the need to exchange/share the same authority data in the networked environment

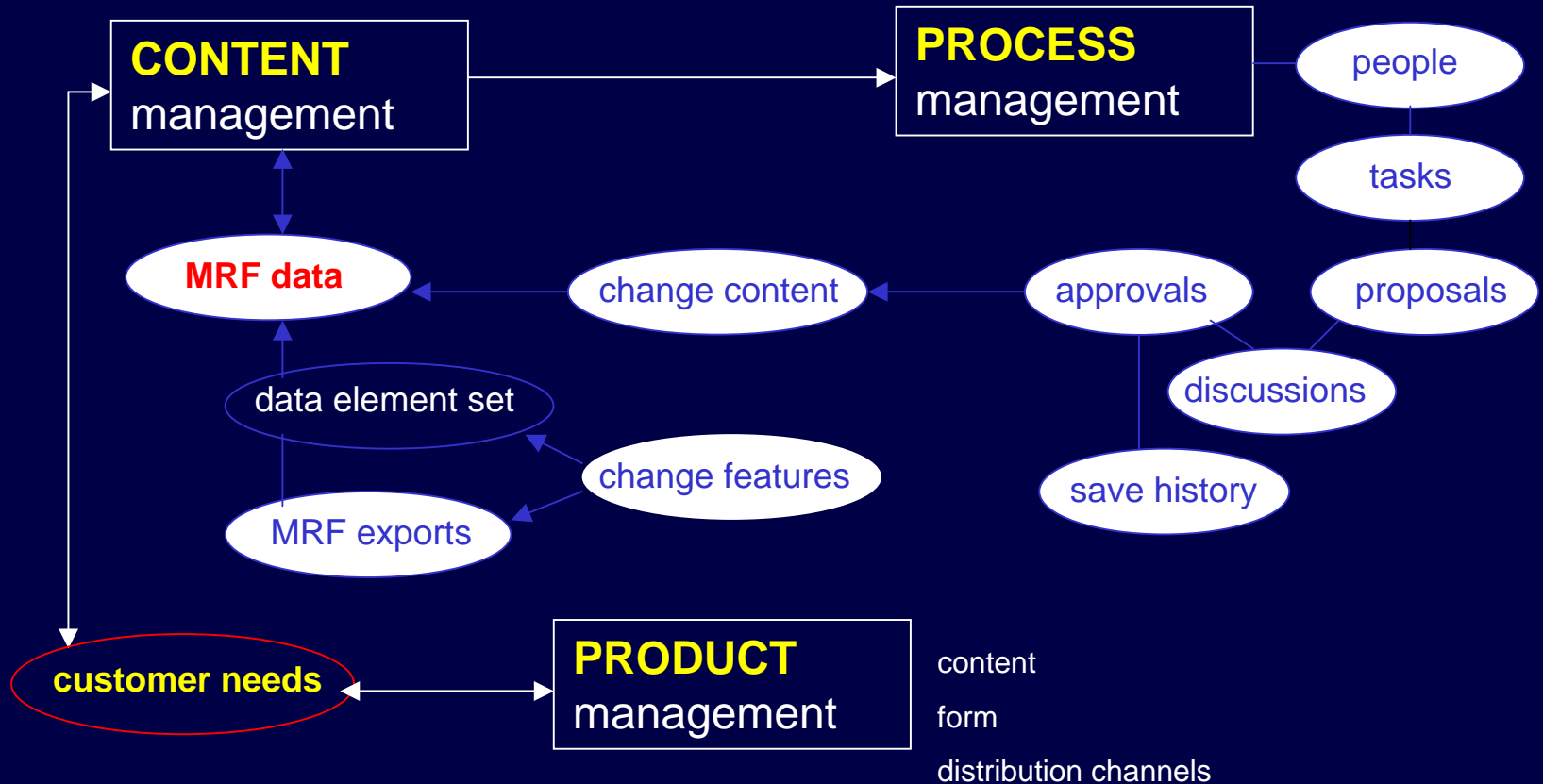
UDC USERS ALSO...

- need UDC data in standard formats for vocabulary exchange (MARC 21/UNIMARC, Topic Maps)
- may prefer XML or XML/RDF binding
- may use UDC behind various kinds of information retrieval systems to support search expansion or vocabulary control
- need multilingual UDC data

UDC MRF DATA UP TO 2006

- CDS/ISIS
- created to support basic maintenance and update and exports needed to produce printed edition at low cost
- text in English (and around 3000 classes in German)
- allowed good text retrieval and export management
- various (but so far, quite proprietary) database exports

2007 - NEW EDITORIAL SYSTEM ...



REASONS

- ensure sustainability of the UDC MRF
 - overcome limitations of CDS/ISIS: locally managed system, data management separate from editorial work
- provide a collaborative system for UDC revisions
 - facilitate inclusion and management of a larger team of internationally distributed editors, manage/track/archive changes
- diversify MRF products
 - expand to MARC formats
 - expand to XML derivatives (of MRF) or MARC
 - expand to ontology formats
 - expand to multilingual editions

ENHANCEMENT OF UDC DATA

- improving UDC MRF format
- correction and consistency checking of MRF content
- new content

IMPROVING UDC MRF DATA FORMAT

1

- structured UDC heading, coding each element of structured notation
- introduce unique identifier of each class (independent of notation) -> URI
- introduce “class history” tracing/linking across editions
- introduce hierarchy data (link to next higher notation)

IMPROVING UDC MRF DATA FORMAT

- link between special auxiliary tables and UDC numbers to which they are applicable
- enable language codes for each textual field
- introduce keyword index element
- introduce mapping elements

IMPROVING UDC EXPORTS

- continue to support old export formats
- providing additional simple textual and XML formats for various purposes e.g.
 - mapping between cancelled/new numbers
 - export on demand (subject selection, hierarchy level selection certain, table selection)
- inclusion of standard exports
 - MARC 21 and UNIMARC authority formats [first]
 - Topic Maps/SKOS [the next step]

NEW UDC MRF DATA

- mapping to other subject systems
 - general classifications (international and national systems)
 - special classifications
 - ? NLM – National Library of Medicine Classification
 - ? SMC – Subject Mathematical Classification
 - ? PACS – Physics & Astronomy Classification Scheme
 - ? INSPEC – physics, engineering
 - thesauri and subject heading systems (AAT, MeSH etc.)
- multilingual data
 - future projects to be negotiated between UDC Consortium members

TIMELINE

- by the end of 2007
 - new editorial system tested
 - data migration to the new database
 - adding data 1st phase

- the first half of 2008 ...
 - adding data 2nd phase
 - XML/MARC exports

- the second half of 2008
 - starting work on exports in ontology formats
 - starting work on adding mappings for certain areas
 - starting work on multilingual data to a certain level

SUMMARY

- UDC is a powerful versatile indexing system that can be used to an arbitrary level of detail
- the use of classification in indexing and retrieval depends on the management of subject data and on the quality of the data
- UDC Consortium is currently working on the enhancement of UDC data that would
 - facilitate creation of UDC authority file
 - facilitate sharing and use of such a file in library systems or information systems alike
 - expand UDC MRF with multilingual data and mappings that would help



Many thanks!