6 Classification Schemes

Fachhochschule Nordwestschweiz
Hochschule für Wirtschaft

Organizing Information

- Objective: finding a way through the overwhelming volume of material.
- Approach: organize information into patterns with related items brought together
- Information collections used by many people are organized in a way that correspond to the needs of most users, e.g.
  - the navigation on the intranet or on a homepage
  - the arrangement of books in libraries or bookshops

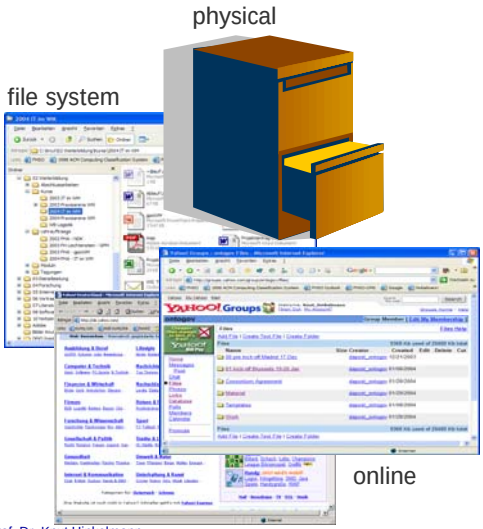
  support users in searching
- Information collections for individual use are also organized
  - the order of paper files reflects the way you normally use them
  - file management on a computer groups items according to their shared characteristics, e.g. the nature of the item (software, document, database etc) the project reference

Classification

- *Classification* is an organization means arranging information items into classes - dividing the *universe of information* into manageable and logical portions.
- A *class* or *category* is a group of concepts that have something in common. This shared property gives the class its identity.
- Classifications may be designed for various purposes like
  - scientific classification
  - classification for information indexing and retrieval
- A class may be further divided into smaller classes (or subclasses), and so on, until no further subdivision is feasible. So classification is likely to be *hierarchic.*

Source: UDConline
(http://www.udconline.net/)

Use of classification schemes



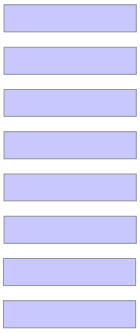
Classification schemes can be used to

- physically group items, e.g.
  - books in a library or
  - retail goods in a supermarket
  - papers in a file cabinet
- logically organize references to information objects - in other words: *metadata* - , e.g.
  - directory on a computer
  - internet directory
  - yellow pages system

Types of classification systems

Flat Organisation:

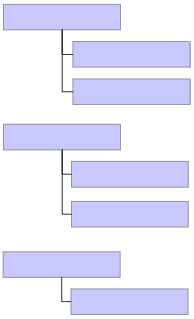
no structure between categories



Taxonomy:

categories arranged in a hierarchical structure

related things are grouped together



Classification Schemes

- The classification scheme can be
  - decided locally
  - represent a consensus

  The greater the quantity or complexity of items, the more helpful it is to follow a ready-made classification scheme, which represents a consensus as to a helpful order of classes
- Classification schemes may be either:
  - special, i.e. limited to a specific domain of interest; or
  - general, i.e. aiming to cover all subjects equally ('the universe of information').
- Three most widely used general classification schemes are :
  - Dewey Decimal Classification (DDC)
  - Universal Decimal Classification (UDC)
  - Library of Congress Classification (LCC)

Example: Universal Decimal Classification (UDC)

- The Universal Decimal Classification (UDC) is a classification scheme for all fields of knowledge and knowledge representation.
- UDC was originally created to organize a universal bibliography
- UDC was created in 1895, it has been translated into over thirty languages
- The scheme is updated annually, its standard version - known as the Master Reference File (MRF) - is available electronically in English language,
- The UDC is structured in a hierarchical manner, based on ten main classes

0 GENERALITIES
1 PHILOSOPHY. PSYCHOLOGY
2 RELIGION. THEOLOGY
3 SOCIAL SCIENCES
4 VACANT
5 NATURAL SCIENCES
6 TECHNOLOGY
7 THE ARTS
8 LANGUAGE. LINGUISTICS. LITERATURE
9 GEOGRAPHY. BIOGRAPHY. HISTORY

The classes are further divided decimally. The notation is basically arabic numerals:

004	*Computer science and technology*
004.8	*Artificial intelligence*
004.89	*Artificial intelligence application systems*
004.891	*Expert systems*
004.891.2	*Consultation expert systems*

http://www.udcc.org/

Applications of UDC

- **libraries**
  - shelf arrangement
  - information retrieval (classified catalogues)
  - collection management (acquisition, circulation statistics, weeding)
- **museums and archives**
  - collection management
  - objects indexing and retrieval
  - collection display
- **bibliographies and bibliographic databases**
  - subject information navigation
  - information retrieval
- **information services**
  - selective dissemination of information (user's profile description)
- **Internet**
  - subject gateways (information presentation and navigation)
  - metadata (information discovery)
- **As a source for building knowledge domain maps** (ontologies), other indexing languages (thesauri) and various kinds of taxonomies and special classifications

Notation

- Most classification schemes, including UDC, have a *notation* - a code that symbolizes the subject of each class and its place in the sequence.
- A simple list of named classes, which would file alphabetically, would not fulfil the purpose of keeping related things together, and separated from unrelated things.
- This can be done by using a notation which has an inherent order, such as numerals, alphabetic notation or a mixture (alphanumeric).
- Notation with variable length can also express the position in the hierarchy, with each extra character representing a lower level; this is called *expressive notation*. Arabic numerals arranged as decimal fractions are ideal for this purpose.
- Decimal fractions also have the advantage of being infinitely extensible, so it is always possible to introduce further subdivisions without altering the ordinal value of the rest of the sequence. Such notation is said to be *hospitable*.

Example 2: Computing Classification Scheme of the Association for Computing Machinery ACM

- Developed to classify articles of the ACM Computing Reviews journal
- In the meanwhile it is used by many other computer science journals, the ACM Digital Library and the MEDOC database
- The full ACM classification scheme involves the following concepts :
  - classification codes: tree structure containing three coded levels
  - subject descriptors: an uncoded fourth level of the tree
  - general terms: predefined set of terms that apply to any elements of the tree

algorithms	languages	security
design	legal aspects	standardization
documentation	management	theory
economics	measurement	verification
experimantation	performance	
human factors	reliability	

  - implicit subject descriptors (also called "Proper Noun Subject Descriptors"): names of products, systems, languages, and prominent people in the computing field, along with the category code under which they are classified

www.acm.org/class/1998

The first three levels of the ACM Classification Scheme

Classification codes

- **A. General Literature**
- **B. Hardware**
- **C. Computer Systems Organisation**
- **D. Software**
- **E. Data**
- **F. Theory of Computation**
- **G. Mathematics of Computing**
- ***H. Information Systems***
- **I. Computer Methodologies**
- ***H. Information Systems***
- **K. Computing Milieux**

H. Information Systems
H.0 General
H.1 Models and Principles
H.2 Database Management
H.3 Information Storage and Retrieval
H.4 Information Systems Applications
H.5 Information Interfaces & Presentation
H.m Miscellaneous

H.3 Information Storage and Retrieval
H.3.0 General
H.3.1 Content Analysis and Indexing
H.3.2 Information Storage
H.3.3 Information Search and Retrieval
H.3.4 System and Software
H.3.5 Online Information Systems
H.3.6 Library Automation
H.3.7 Digital Libraries
H.3.m Miscellaneous

Examples for Subject Descriptors of the ACM Classification Schemes (Level 4 - uncoded)

H.3.1 Content Analysis and Indexing
- Abstracting methods
- Dictionaries
- Indexing methods
- Linguistic processing
- Thesauruses

H.3.3 Information Storage and Retrieval
- Clustering
- Information filtering
- Query formulation
- Relevance feedback
- Retrieval models
- Search processes
- Selection processes

I.2.8 Problem Solving, Control Methods, and Search
- Backtracking
- Control Theory
- Dynamic Programming
- Graph and tree search strategies
- Heuristic methods
- Plan execution, formation, and generation
- Scheduling

Example Classification according to ACM Classification Scheme

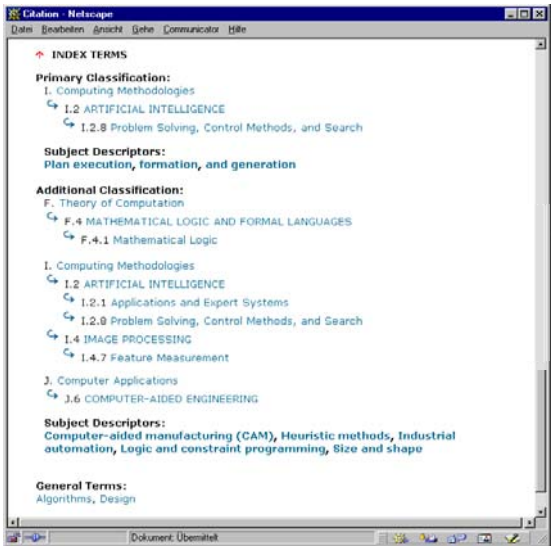
(from ACM Digital Library)

A Consequence-finding approach for feature recognition in CAPP

Knut Hinkelmann

Proceedings of the seventh international conference on Industrial and Engineering applications of artificial intelligence and expert systems May 1994

The document has multiple classifications
- Primary Classification
- Additional Classifications



Problems with Classification Schemes

- Classification Schemes must be revised and adapted to new developments
  - Example: The ACM classification scheme
- Classification Schemes often are developed for a specific domain of interest. Documents outside the scope cannot be classified.
- Classification scheme must be comprehensible for all users
- Many documents cannot be classified unambiguously. To deal with this problem, classification systems can offer different solutions
  - Select exactly one classification
    - Example: Physically organising books in a library
  - Assign multiple classifications
  - Assign one primary and optional additional classifications
    - Example: In a library the primary classification corresponds to the physical location while the additional classifications can be used for searching in the library catalogue

Monodimensional vs. Polydimensional Classification

Example

Dokument
- [Aspect: Type]
  - report
  - correspondence
  - invoice
  - presentation
- [Aspect: kind]
  - text
  - image
  - video
- [Aspect: Format]
  - ASCII
  - XML
  - doc
  - ppt
  - txt

- Monodimensional
  - Classification according to one aspect
- Polydimensional
  - Classification according to different (independent) aspects
  - Documents may have different classes
  - Each dimension corresponds to one aspect/attribute

„Foliensatz Information Retrieval“

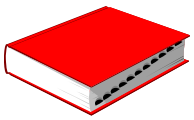
Example: Polydimensional Classification

- Example: Document management in a re-insurance company
- Documents are classified in two dimension
  - Product type
  - Market in which the product is offered

Market:
- Europe
  - Switzerland
  - EU
  - Eastern Europe
- North America
  - USA
  - Canada
- South America
- Asia
  - China
  - India
  - Japan
- Africa

Products:
- health insurance
  - in-patient
  - out-patient
  - dental
  - short-term disability
- life insurance
  - endowment
  - term
  - accident
- property insurance
  - household
  - liability
    - car
    - private

Each classification dimension corresponds to a metadata attribute



Index:

Document type:	report
Document format:	MS Word
Product:	Life - annuity
Market:	Europa - Switzerland