# Managing classification in libraries: a methodological outline for evaluating automatic subject indexing and classification in Swedish library catalogues

Koraljka Golub, Joacim Hansson,
Dagobert Soergel, Douglas Tudhope

UDC-seminar, Lisbon 30 october 2015

# Points of departure

- (Semi)automated subject indexing and classification represents a potential solution to retain established objectives of bibliographic systems in digital document enivornments.

- The Swedish library system is interesting in that it has recentely implemented DDC at a national level, and research in this new environment is scarce.

# Project objectives

- The aim of this project is to gain a scientifically sound understanding of the level to which it is possible to apply automatic subject indexing to Swedish textual resources, specificaly to assign DDC classes (major focus) and Swedish Subject Headings (SAO) (minor focus), as well as indexing with free keywords derived from the resource at hand.

- Objectives:
  - Establish a gold standard for DDC, SAO and free keywords,
  - determine an inter-indexer consistency level between cataloguers, end users, subject experts and automatic output,
  - compare (several) automatic indexing applications against the gold standard,
  - determine what terms lead to best end-user retrieval results
  - analyse differences in the above between the domains of applied sciences and the humanities.



# Methodology

- Test collection
- Automatic indexing tools
- Gold standard
- Retrieval test
- Domain analysis



$$\text{Indexing recall} = \frac{\text{\# of correct index terms assigned by the tool}}{\text{\# of correct index terms in the giold standard}}$$

$$\text{Indexing precesion} = \frac{\text{\# correct index terms assigned by the tool}}{\text{\# of all index terms assigned by the tool}}$$



# Methodology

-Test collection
-Automatic indexing tools
-Gold standard
-Retrieval test
-Domain analysis

# Domain analysis

Different **knowledge domains** have different knowledge structures and knowledge claims relating to the division of labour in society:

- ”Pure sciences”
- Applied sciences
- Social Sciences
- Humanities

Terminological conventions
Epistemological assumptions
Relation to social development

Other domain relevant aspects:

- Geographical context
- Contingency (the emergence of new disciplinary structures, e.g mutidiciplinary researh fields)
- Document environments

# Domain analysis

Apart from knowledge domains, Domain Analysis deals with power distribution:

Knowledge domain ⟶ Power domain

**Social**: end-users, cataloguers, subject experts

**Institutional**: automated indexing/classification tools, document environment – the DDC itself.

**Cognitive**: end-users, cataloguers, subject experts – relates to cognitive authority (relevant when deciding on the gold standard)

# Significance of study

Value comparison of the creation of subject index terms between

- professional
- end-user
- automatic tools

Study proposes way to deal with a number os issues, like:

- cheap assignment of controlled subject terms
- hierarchical browsing by subject
- retrievability of Swedish resources in multilingual systems and their integration into the Semantic Web.

Combining different kinds of methodological and theoretical approaches in new ways.



# Thank you.

Joacim.hansson@lnu.se