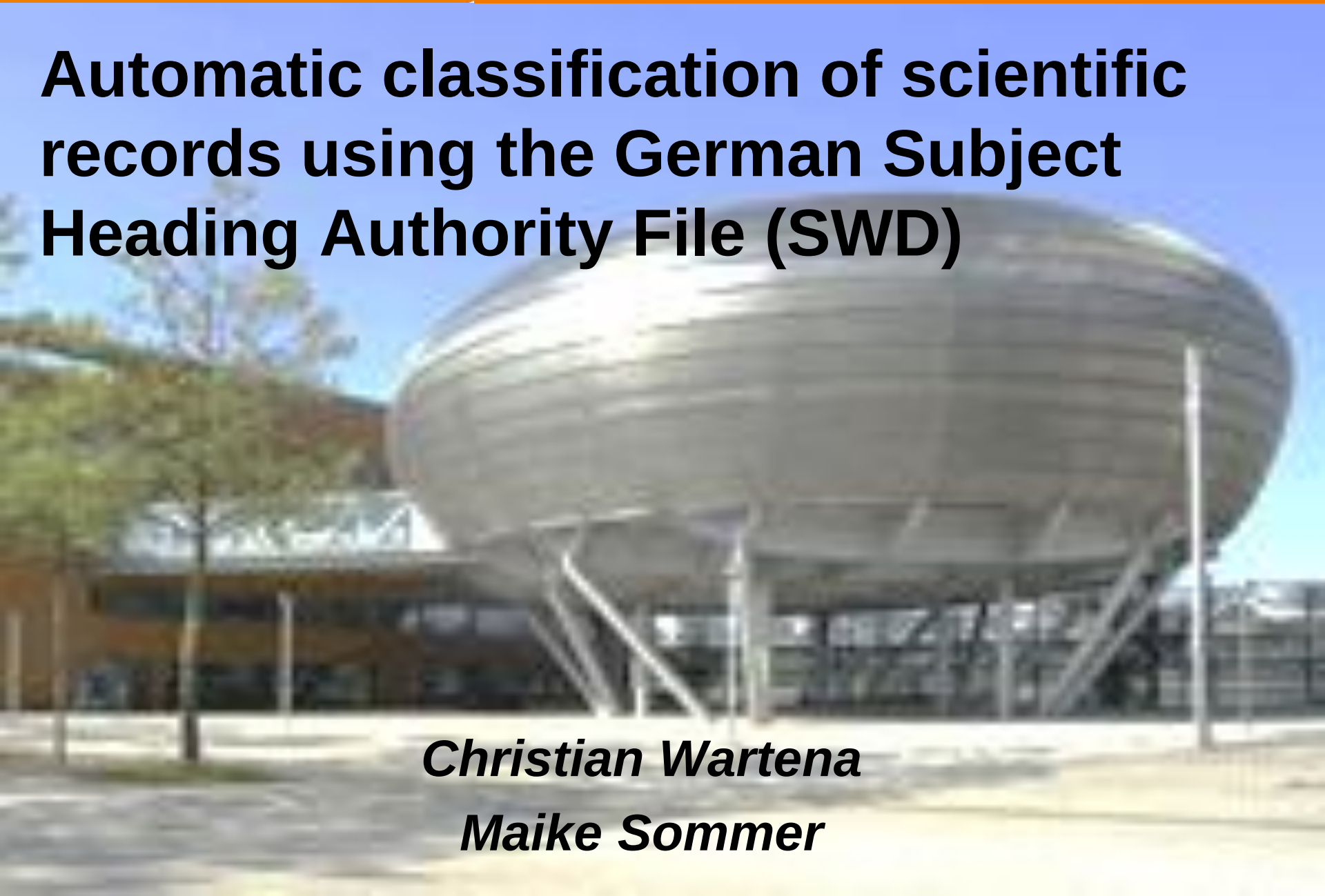
Hochschule Hannover
University of Applied Sciences and Arts

Fakultät III – Medien, Information und Design

Automatic classification of scientific records using the German Subject Heading Authority File (SWD)



Christian Wartena

Maike Sommer

Introduction

- Simple methods can work very well.
- Traditional thesauri can be extremely useful for automatic classification.

Overview

- Problem
- Context
- Approach
- Experimental Results

Problem

- University libraries archive and publish works written by staff and students.
  - Some scientific articles
  - Master Thesis, PhD-Thesis
  - Speeches
  - Reports, Project reports
- Publications need to be classified to support search and browsing.

Problem

- Authors don‘t have knowledge of the classification system
  - Different levels of granularity
  - Inconsistency
  - Many synonyms used for one concept.
- Librarians
  - don‘t have time
  - don‘t have detailed knowledge of all disciplines
- Human Classification
  - Will become inconsistent over time.

Overview

- Problem
- Context
- Approach
- Experimental Results

Dewey Decimal Classification

- Thesaurus covering all scientific disciplines
- > 50 000 Classes, Organized hierarchically.
- Used Worldwide; Many translations
- Kept up-to-date by Library of Congress

DDC Highest levels

000 Computer science, information & general works
100 Philosophy & psychology
200 Religion
300 Social sciences
400 Language
500 Science
600 Technology
700 Arts & recreation
800 Literature
900 History & geography

DDC Example

„Participatory Democracy in France“ **323.**0420944

300 Social sciences

320 Political science

323 Civil & political rights

Auxiliary table 1: Special topics

Auxiliary table 1: Historical, geographic, persons treatment

Auxiliary table 2: France and Monaco

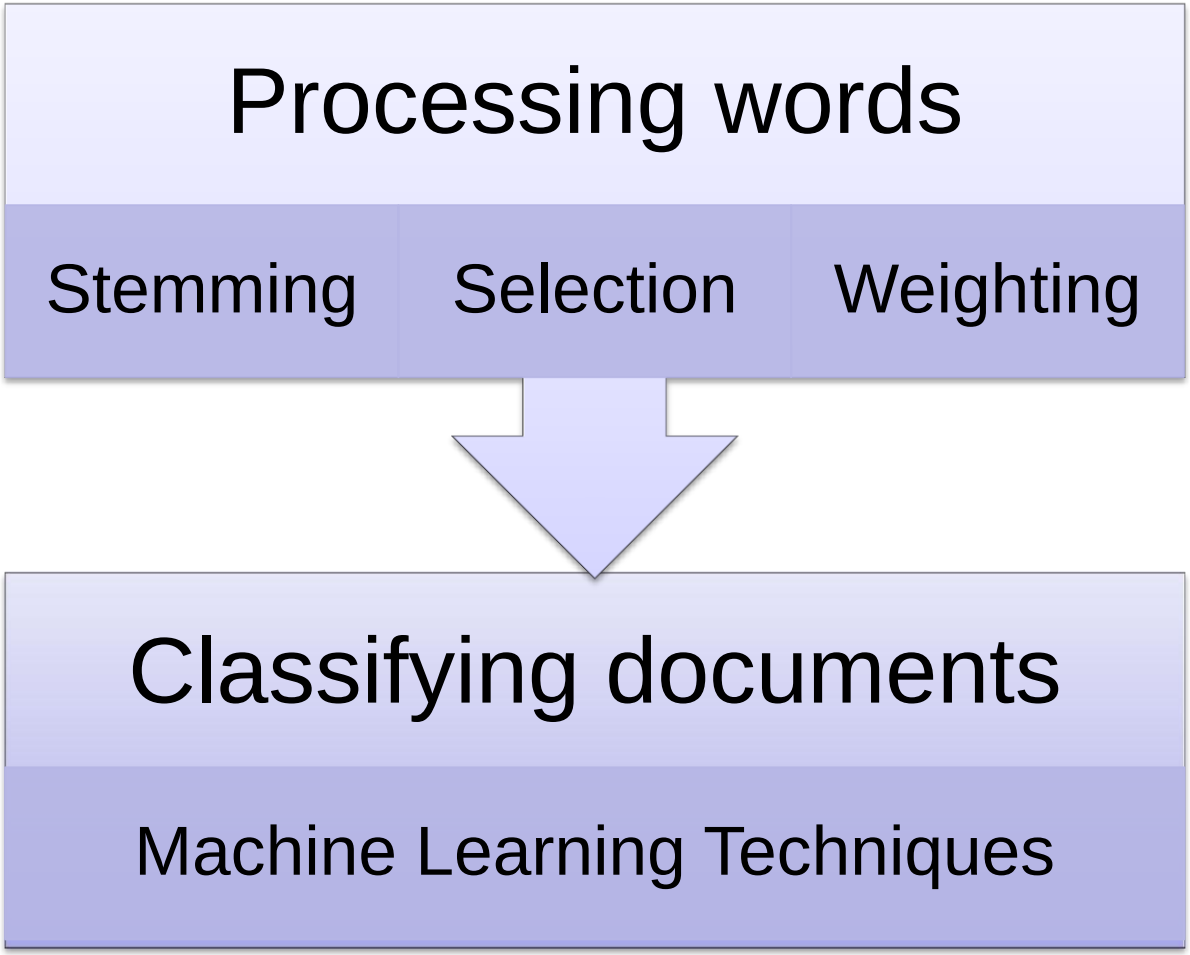
DDC Subject Areas

- Many (German) libraries use the second level of the DDC-Hierarchy as Subject Areas (000, 010, 020 etc.)
  - 100 subject areas
- Inconsistency w.r.t. Computer Science:
  - Some libraries use 000 (General Science)
  - Others use 004

SWD

- Schlagwortnormdatei = German Subject Authority File
  - Maintained by German National Library
  - ~188 000 descriptors
  - ~155 000 non-descriptors
- Links from (almost all) SWD-Subject Headings to DDC Classes.
  - Project of German National Library and UAS Cologne (Criss Cross)

Text-based Classification



DDC Classification of Scientific Records

- DFG-funded project *Automatic Enhancement of OAI Metadata.*
  - Automatic Classification Toolbox for Digital Libraries.
  - English and German Texts

Overview

- Problem
- Context
- Approach
- Experimental Results

Our Approach

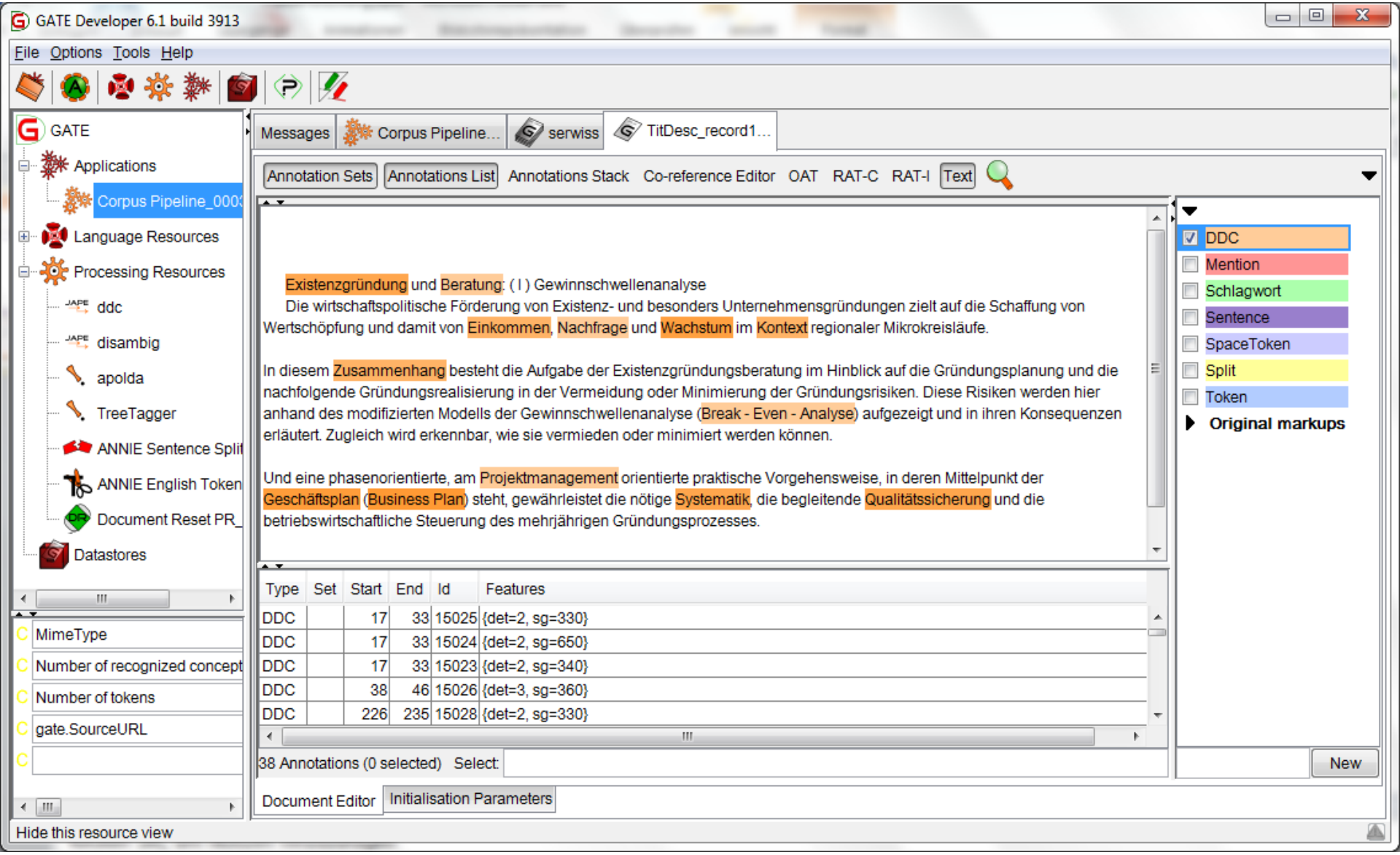
- Don’t use a corpus
  - No training on documents
  - No idf weighting
  - No bias
  - Suited for small repositories
- Everything is done in the first phase
  - Classification is almost trivial
  - Search SWD terms in abstracts
  - Use links to DDC for classification

Preprocessing

- Lemmatization
  - Tree Tagger
  - Not just Porter stemmer!
- POS-Tagging
- Thesaurus-Lookup
  - Subject headings are not words as found in running texts!
  - Some terms have been removed.
  - Terms have to match, single terms have to be nouns
  - Ambiguous terms are removed

Ambiguity

- A **term** *t* is unambiguous if
  - *t* is the preferred label of exactly one concept (regardless of non-preferred labels)
  - *t* is the alternative label of exactly one concept and doesn't occur as preferred label.
- A **concept** might be connected with several DDC-classes.
  - That is something else.


GATE Developer 6.1 build 3913
File Options Tools Help
GATE
Applications
Corpus Pipeline_0003
Language Resources
Processing Resources
ddc
disambig
apolda
TreeTagger
ANNIE Sentence Split
ANNIE English Token
Document Reset PR_
Datastores
MimeType
Number of recognized concept
Number of tokens
gate.SourceURL
Messages
Corpus Pipeline...
serwiss
TitDesc_record1...
Annotation Sets
Annotations List
Annotations Stack
Co-reference Editor
OAT
RAT-C
RAT-I
Text
Existenzgründung und Beratung: (I) Gewinnschwellenanalyse
Die wirtschaftspolitische Förderung von Existenz- und besonders Unternehmensgründungen zielt auf die Schaffung von Wertschöpfung und damit von Einkommen, Nachfrage und Wachstum im Kontext regionaler Mikrokreisläufe.
In diesem Zusammenhang besteht die Aufgabe der Existenzgründungsberatung im Hinblick auf die Gründungsplanung und die nachfolgende Gründungsrealisierung in der Vermeidung oder Minimierung der Gründungsrisiken. Diese Risiken werden hier anhand des modifizierten Modells der Gewinnschwellenanalyse (Break - Even - Analyse) aufgezeigt und in ihren Konsequenzen erläutert. Zugleich wird erkennbar, wie sie vermieden oder minimiert werden können.
Und eine phasenorientierte, am Projektmanagement orientierte praktische Vorgehensweise, in deren Mittelpunkt der Geschäftsplan (Business Plan) steht, gewährleistet die nötige Systematik, die begleitende Qualitätssicherung und die betriebswirtschaftliche Steuerung des mehrjährigen Gründungsprozesses.
DDC
Mention
Schlagwort
Sentence
SpaceToken
Split
Token
Original markups
Type Set Start End Id Features
DDC 17 33 15025 {det=2, sg=330}
DDC 17 33 15024 {det=2, sg=650}
DDC 17 33 15023 {det=2, sg=340}
DDC 38 46 15026 {det=3, sg=360}
DDC 226 235 15028 {det=2, sg=330}
38 Annotations (0 selected) Select:
New
Document Editor
Initialisation Parameters
Hide this resource view

Classification

- A subject heading might have links to several DDC classes
  - with confidence levels 1 to 4
  - Let *ddc(t)* be the set of all classes a term *t* is linked to **with confidence level > 1**.
- If *c* is a class, c^n is the broader class of *c* at the n^{th} level.
- The contribution of a term *t* to a class *c* now is:
  - $w(t,c) = \frac{|\{c_i \in ddc(t) | c_i^n = c\}|}{|\{c_i \in ddc(t)|}$
- Sum up these values over all recognized terms in a text:
  - $w(T,c) = \Sigma_{t \in \mathrm{T}} w(t,c)$

Overview

- Problem
- Context
- Approach
- Experimental Results

Experiment

- OAI Metadata from 7 German University Libraries
  - Retrieved title, abstract, keywords, publication type and DDC subject area for all publications in German.
  - No language tags for abstracts, but 1st retrieved abstract is always the German one.
  - > 3500 records
- All records classified
  - ACT-DL web service
  - Our method

Evaluation

- Evaluation on
  - 1st DDC Level (10 Categories)
  - 2nd DDC Level (100 Categories)
- Both methods give ranked lists
  - Precision of top 5 results (prec@5)
  - Mean Reciprokal Rank

Results /1

- Different sources of information
- Our own library

DDC Level	Sources	SWD-Based		ACT-DL	
		MRR	Rec@5	MRR	Rec@5
1	Title + abstract	0,68	0,89	0,67	0,90
2	Title + abstract	0,48	0,66	0,39	0,37
2	Title + keywords	0,61	0,76	0,32	0,39
2	Title + abstract + keywords	0,61	0,77	0,39	0,47

Results /2 (Title + Abstract)

		SWD-Based		ACT-DL	
Repository	Size	MRR	Rec@5	MRR	Rec@5
Hanover UAS	271	**0,48**	**0,63**	0,39	0,37
Cologne UAS	254	0,32	**0,44**	0,35	0,39
Frankfurt UAS	120	**0,55**	**0,75**	0,49	0,62
TU Berlin	2036	0,41	0,61	0,59	0,66
Uni. Hildesheim	97	0,25	**0,39**	0,25	0,29
Uni. Regensburg	790	0,61	**0,80**	0,65	0,72
Freiburg UE	258	**0,53**	**0,75**	0,33	0,36

Results /3

		SWD-Based		ACT-DL	
Repository	Size	MRR	Rec@5	MRR	Rec@5
PhD thesis	2503	0,47	0,66	0,63	0,70
Master thesis	277	0,32	**0,44**	0,37	0,41
Essay	195	**0,59**	**0,75**	0,29	0,36
Monograph	176	**0,46**	**0,62**	0,43	0,51
Festschrift	106	**0,43**	**0,76**	0,38	0,40
Lecture	84	**0,40**	**0,96**	0,03	0,06

Conclusions 1

- Usefulness of SWD
  - Subject heading file can be used as a lexical thesaurus!
  - Links to DDC are very useful.
- Method
  - Simple classification method not as good as advanced ML approach,
  - but when data diverge from training data (or if no training data available) the proposed method performs equal well or even better.
  - Methods like this one can serve as a baseline.

Conclusions 2

- Simple methods can work very well.
  - But don't use too simple methods.
  - Using a massive amount of hand-crafted classifications
- Traditional Thesauri can be extremely useful for automatic classification.
  - We did of course not use the full potential of the SWD.

Thanks for your attention.

- Questions?