# Automatische Klassifikation mit linguistischen Methoden

Mathias Lösch

30. Oktober 2009

1 Motivation

2 Ist-Situation

3 Projektziele

4 Automatische Klassifikation

5 Ausblick & Zusammenfassung

# Motivation: Was wir gerne hätten . . .

- Semantische Suche: Suche nach Konzepten unabhängig von der sprachlichen Repräsentation (z. T. schon realisiert durch Anfrageerweiterung mit EUROVOC bzw. Indexerweiterung mit SWD)
- Browsing nach Fachgebieten
- Semantisches Browsing: Assoziationen zwischen ähnlichen Dokumenten; sinntragende Hyperlinks zwischen Dokumenten, z. B.: *konkretere Dokumente*, *abstraktere Dokumente* (Kuhlen, 2002)

# Ist-Situation: Was dem im Wege steht . . .

Derzeitige Sacherschließung der OAI-Records über *Dublin Core*

- ist zu primitiv:
  - beim Mapping von detaillierteren Formaten geht Information verloren
  - Felder können ohne weitere Qualifizierung mehrfach vergeben werden
- kennt keine entsprechenden Katalogisierungsregeln, nur Hinweise für die Ausfüllung der Felder (Eversberg, 2008; Tennant, 2004)

# Beispiel: Informationen in `dc:subject` (Pieper u. Summann, 2006)

Empfehlung: Stichwörter, Schlagwörter oder Notationen
Praxis:

- Schlagwörter, Deskriptoren
- Klassifikationsnotationen, z.B.:
  - Basisklassifikation, MSC, LoC, PACS, DDC
- Klassenbezeichner
- Wiederholung von Autoren, Titel, etc.

Lösungsansätze:

- Standardisierung: DINI-Zertifikat, DRIVER Guidelines
- Automatische Homogenisierung der Sacherschließung

# Projektziele

- Automatische Homogenisierung der OAI-Sacherschließung für den BASE-Index durch automatische Dokumentenklassifikation nach Dewey Decimal Classification (DDC)
- Integration der neuen Sacherschließungsinformationen in den BASE-Index und in die BASE-Benutzerschnittstelle (Browsing)
- Nachnutzung der Daten durch andere Organisationen (Services, inhaltsorientierte Vernetzung von Repositorien)

# Automatische Klassifikation: Dewey Decimal Classification

## Haupttafeln (Schedules)

| | |
|---|---|
| 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 |

# Automatische Klassifikation: DDC

## Notationsstruktur

600 Technik und Technologie
630 Landwirtschaft und verwandte Bereiche
636 Viehwirtschaft
636.7 Hunde
636.728 Pudel

# Notationssynthese

»Presidential system in Hawaii« 321.804209969

# Notationssynthese

»Presidential system in Hawaii« 321.804209969

Social sciences

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states
Democracy

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states
Democracy
Hilfstafel 1: Standardschlüssel 042: Spezielle Themen

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states
Democracy
Hilfstafel 1: Standardschlüssel 042: Spezielle Themen
Hilfstafel 1: Standardschlüssel 09: Historische, geografische, personenbezogene Behandlung

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states
Democracy
Hilfstafel 1: Standardschlüssel 042: Spezielle Themen
Hilfstafel 1: Standardschlüssel 09: Historische, geografische, personenbezogene Behandlung
Hilfstafel 2: Geografische Gebiete, Zeitabschnitte, Personen: Andere Teile der Welt und außerirdische Welten

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states
Democracy
Hilfstafel 1: Standardschlüssel 042: Spezielle Themen
Hilfstafel 1: Standardschlüssel 09: Historische, geografische, personenbezogene Behandlung
Hilfstafel 2: Geografische Gebiete, Zeitabschnitte, Personen: Andere Teile der Welt und außerirdische Welten
Andere Teile des Pazifischen Ozeans Polynesien

# Notationssynthese

»Presidential system in Hawaii« 321.804209969
Social sciences
Political science
Systems of governments & states
Democracy
Hilfstafel 1: Standardschlüssel 042: Spezielle Themen
Hilfstafel 1: Standardschlüssel 09: Historische, geografische, personenbezogene Behandlung
Hilfstafel 2: Geografische Gebiete, Zeitabschnitte, Personen:
Andere Teile der Welt und außerirdische Welten
Andere Teile des Pazifischen Ozeans Polynesien
Nordmittelpazifische Inseln Hawaii

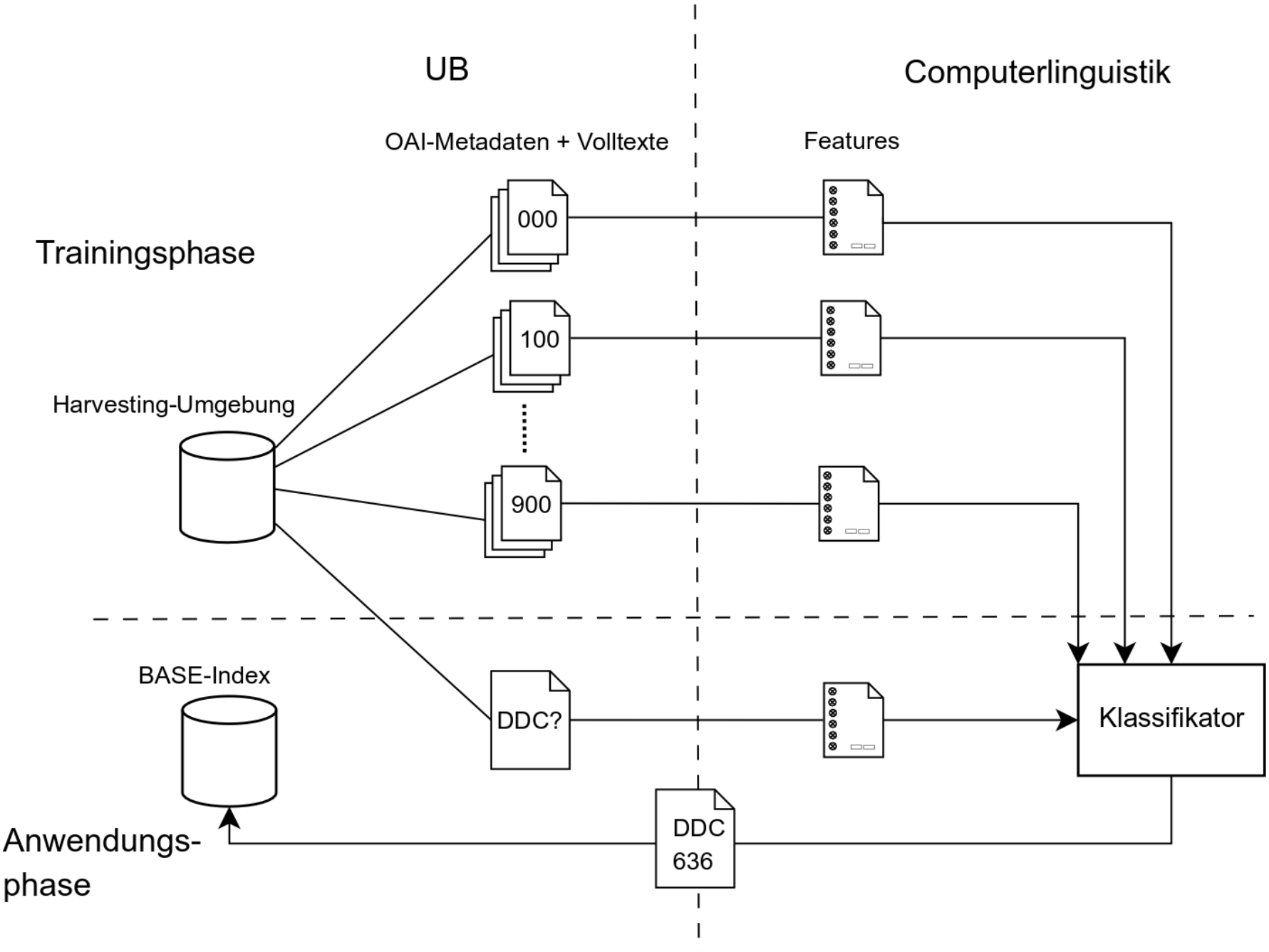
# DDC: Vor- und Nachteile (nach Bertram, 2005, S. 196f)

Vorteile:

- Universalität
- Sprachunabhängigkeit durch numerische Notation
- international am weitesten verbreitete Klassifikation
- zunehmende Verwendung in Europa, durch Verwendung in der DNB auch in Deutschland

Nachteile:

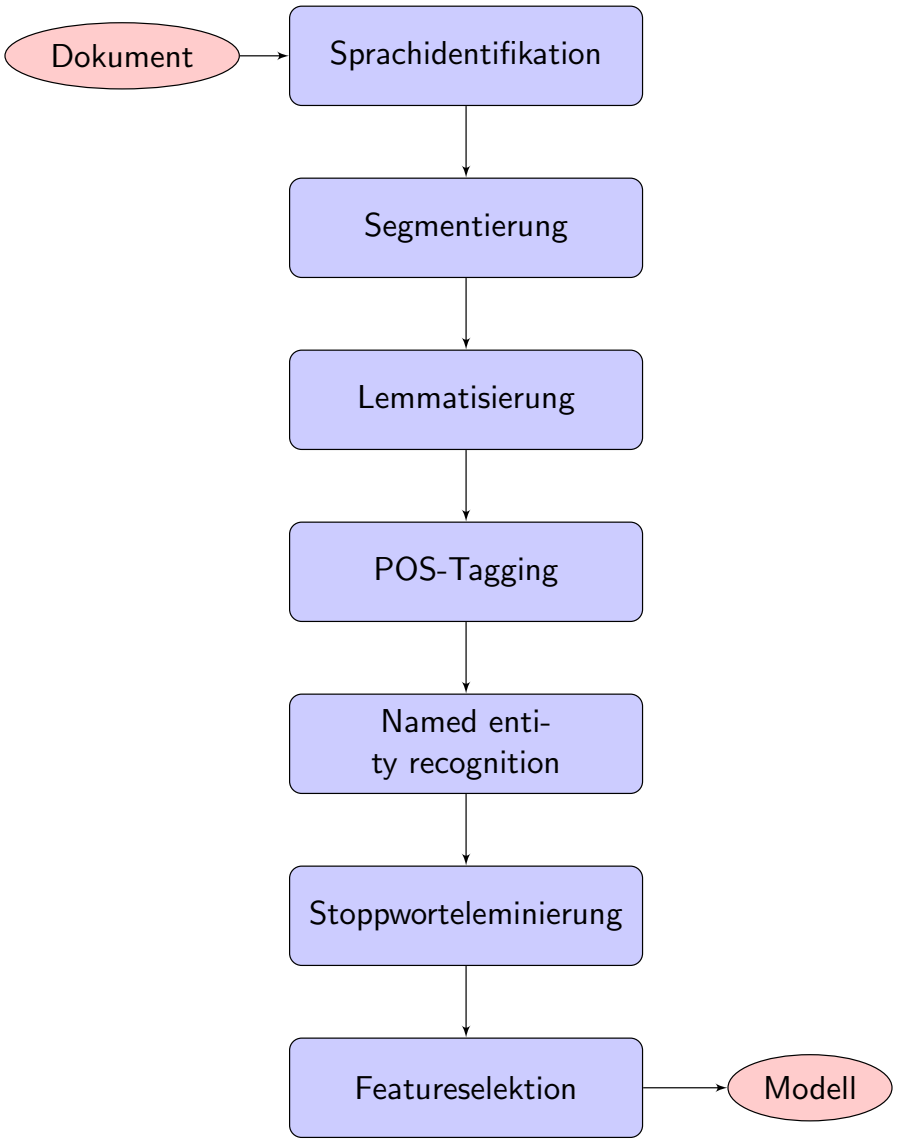
- stark angloamerikanisch geprägt
- Mängel in der Hierarchie
- Notationssynthese erfordert viel Erfahrung und Sachkenntnis


UB
Computerlinguistik
OAI-Metadaten + Volltexte
Features
Trainingsphase
000
100
900
Harvesting-Umgebung
BASE-Index
DDC?
Klassifikator
DDC
636
Anwendungs-
phase

# Gewinnung der Trainingsdaten

- Trainingsdaten: OAI-Records mit DDC-Annotationen
  - `dc:title`, `dc:description`, `dc:subject` (Mehler u. Waltinger, 2009) + Volltext
- Gewinnung:
  - falls DDC-Notation vorhanden: direkte Übernahme aus OAI-Records
  - sonst: Ableitung der DDC-Klassen aus anders klassifizierten OAI-Records durch Konkordanzen: MSC, LoC, BK ...

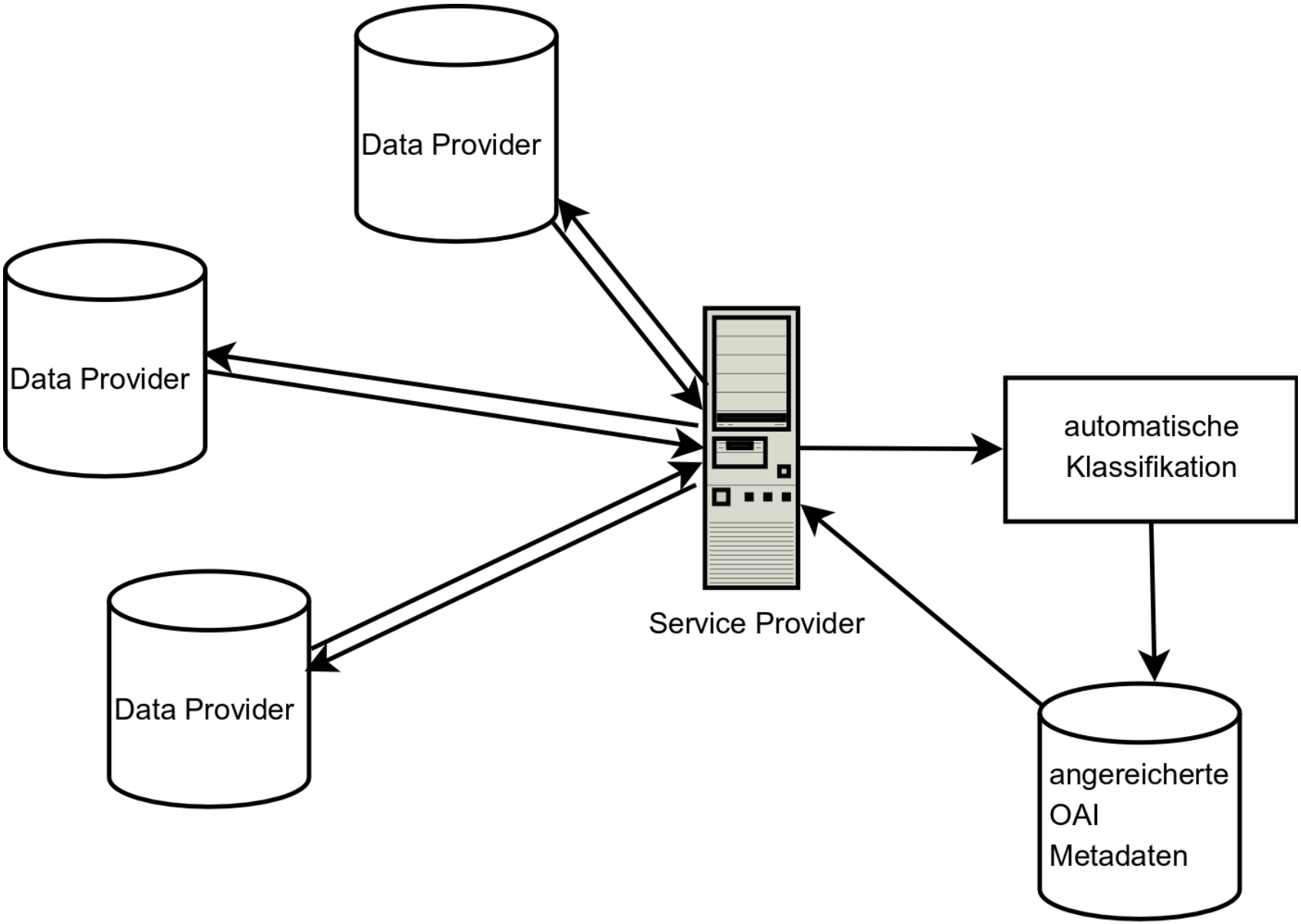
# Feature extraction (Mehler u. Waltinger, 2009)



# Klassifikationsmethoden

- maschinelles Lernen: Support Vector Machines (SVMs) (Joachims, 1998) liefern derzeit die besten Ergebnisse
- Suchmaschinen- bzw. wikipediabasierte Klassifikation als Alternativen (Mehler u. Waltinger, 2009)

# Nachnutzung durch andere Organisationen



# Ausblick: Zukünftige Aktivitäten + Meilensteine

- Struktursensitive Dokumentklassifikation
- Evaluation der Klassifikationsergebnisse
- Integration der Klassifikationsergebnisse in den BASE-Index
- Nutzbarmachung der angereicherten Metadaten über eine Benutzerschnittstelle (Webbasiert, Browser-Plugin)
- inhaltsorientierte Vernetzung von Repositorien
- Evaluation unter Produktionsbedingungen

Meilensteine

1. Nach $^1/_2$ Jahr: aus Trainingsdaten gelernte Modelle + Korpus von zu klassifizierenden Texten
2. Nach 1 Jahr: Grundlegende Funktionen der automatischen Klassifikation
3. nach 1 $^1/_2$ Jahren: Produktivbetrieb

# Zusammenfassung

Es geht um

- Verbesserung der OAI-Sacherschließungsinformationen
- durch automatische Klassifikation nach DDC
- und Nutzung der angereicherten Daten in BASE sowie die Zurverfügungstellung an andere Institutionen.

# Literatur I

[Bertram 2005] Bertram, Jutta: *Einführung in die Inhaltserschließung. Grundlagen – Methoden – Instrumente*. Würzburg : Ergon, 2005

[Eversberg 2008] Eversberg, Bernhard: *Dublin Core*. `http://www.allegro-c.de/formate/kap107.htm`. Version: 2008

[Joachims 1998] Joachims, Thorsten: Text categorization with Support Vector Machines: Learning with many relevant features. In: *Machine Learning: ECML-98*. Springer, 1998, S. 137–142

[Kuhlen 2002] Kuhlen, Rainer: *Hypertextifizierung - Zu den methodischen Grundlagen nicht- linear organisierter Informationssysteme: Text – Kontext – Hypertext*. `http://www.inf-wiss.uni-konstanz.de/CURR/summer02/hypertext/methodische_grundlagen_htx.pdf`. Version: 2002

[Mehler u. Waltinger 2009] Mehler, Alexander ; Waltinger, Ulli: Enhancing Document Modeling by Means of Open Topic Models: Crossing the Frontier of Classification Schemes in Digital Libraries by Example of the DDC. In: *Appears in Library Hi Tech* (2009)

# Literatur II

[Pieper u. Summann 2006] Pieper, Dirk ; Summann, Friedrich: Bielefeld Academic Search Engine (BASE): an end-user oriented institutional repository search service. In: *Library Hi Tech* 24 (2006), Nr. 4, 614–619. `http://eprints.rclis.org/9160/`

[Tennant 2004] Tennant, Roy: Digital Libraries: Metadata's Bitter Harvest. In: *Library Journal* (2004), Nr. 12. `http://www.libraryjournal.com/article/CA434443.html`