

# Automatische Klassifikation in der Praxis

Mathias Lösch

Universitätsbibliothek Bielefeld

[Mathias.Loesch@uni-bielefeld.de](mailto:Mathias.Loesch@uni-bielefeld.de)

Kolloquium Wissensinfrastruktur WS 2011/12

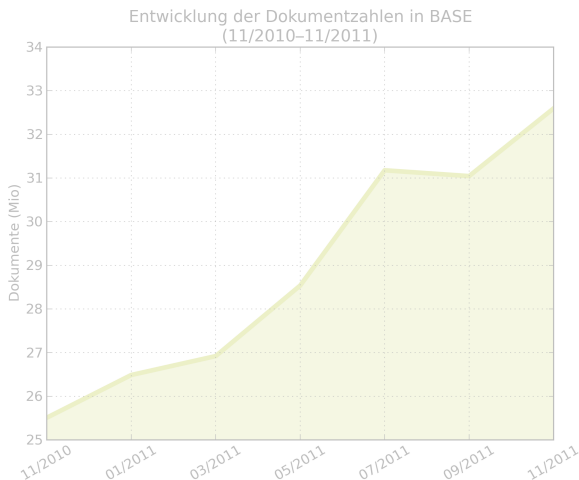
11/11/11

- 1 Motivation
- 2 Wie funktioniert automatische Sacherschließung?
- 3 Praxisbeispiele UB Bielefeld
- 4 Zusammenfassung

- 1 **Motivation**
- 2 Wie funktioniert automatische Sacherschließung?
- 3 Praxisbeispiele UB Bielefeld
- 4 Zusammenfassung

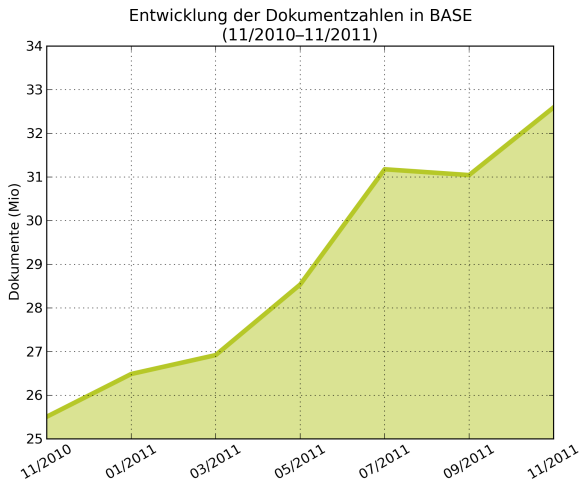
# Motivation: Information Overload

- Starke Zunahme elektronischer Publikationen
- Beispiel BASE: Zunahme nachgewiesener Dokumente seit November 2010 um ~7 Mio.
- ~19.000 Dokumente pro Tag



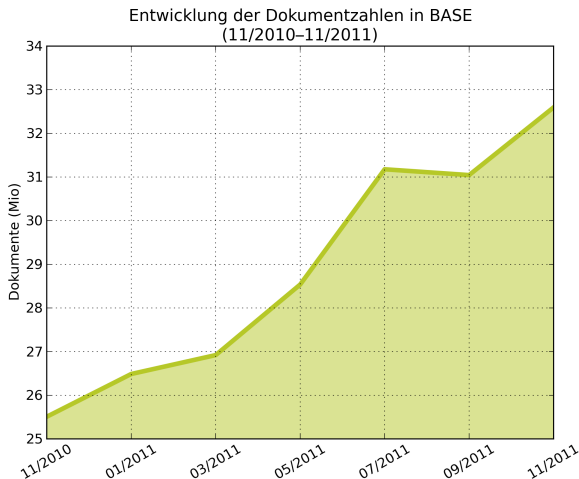
# Motivation: Information Overload

- Starke Zunahme elektronischer Publikationen
- Beispiel BASE: Zunahme nachgewiesener Dokumente seit November 2010 um ~7 Mio.
- ~19.000 Dokumente pro Tag



# Motivation: Information Overload

- Starke Zunahme elektronischer Publikationen
- Beispiel BASE: Zunahme nachgewiesener Dokumente seit November 2010 um ~7 Mio.
- ~19.000 Dokumente pro Tag



# Automatische Sacherschließung im Bibliothekskontext

UL Michigan: Topic Models (Hagedorn et al., 2007)

DLF OAI Portal Search Results

DLF  
DIGITAL LIBRARY FEDERATION

Home Search Browse About Contributors Help

You found **155015** records from **Art and Design**.  
Try a [search](#) to retrieve a different set of records.

Browse by Topic **Browse by Data Contributor**

**Arts & Humanities (1014093)**

- American and Canadian Studies (183681)
- Anthropology and Archaeology (97395)
- Architecture (300864)
- Art History (86550)
- Art and Design (155015)**
- East Asian Languages and Cultures (34461)
- English Language and Literature (125712)
- History (General) (383174)
- Humanities (General) (10671)
- Linguistics (10398)
- Music and Dance (54862)
- Philosophy (41549)
- Religious Studies (7217)
- Russian and East European Studies (36714)

Jump to Records: **1** | 11 | 21 ... 15501 ... 31001 ... 46501 ... 62001 ... 77501 ... 93001 ... 108501 ... 124011 ... 139511 ... 155011

**Record 1 of 155015**  
[add to bookbag](#)

Title	Stray Pieces of Early Christian Writing
Author/Creator	Sherman E. Johnson
Publisher	University of Chicago Press
Year	2003-01-13T16:22:49Z
Resource Type	journal article
Resource Format	application/pdf
Language	English
Source	Journal of Near Eastern Studies, Volume 5, Page 40
Note	10.1086/370769
Subject	epstein-barr virus; weaving; historical collection; voltage
Subject	Art and Design; Internal Medicine and Specialties; Electrical Engineering; History (General)
Subject	Science; Arts & Humanities; Engineering; Social Sciences; Health Sciences
URL	<a href="http://www.journals.uchicago.edu/cgi-bin/resolve?rid=doi:10.1086/370769">http://www.journals.uchicago.edu/cgi-bin/resolve?rid=doi:10.1086/370769</a>
Data Contributor	The University of Chicago Press Journals Division

**Record 2 of 155015**  
[add to bookbag](#)

Title	The Artist of the Egyptian Old Kingdom
Author/Creator	John A. Wilson
Publisher	University of Chicago Press

- Automatische Zuordnung von Datensätzen zu Themen (Topics)

# Automatische Sacherschließung im Bibliothekskontext

Deutsche Nationalbibliothek: Projekt PETRUS (Schöning-Walter, 2010)



- Automatische Verschlagwortung von Netzpublikationen mit SWD
- Automatische Sachgruppenvergabe (DNB-Sachgruppen) für Netzpublikationen



# Automatische Sacherschließung im Bibliothekskontext

UB Bielefeld: DFG-Projekt »Automatische Anreicherung von OAI-Metadaten«

The screenshot shows the BASE web interface with the URL <http://base-search.net/Browse/Dewey>. The page title is "BASE (Bielefeld Academic Search Engine): Browse the Collection". The navigation menu includes "Basic Search", "Advanced Search", "Browsing", and "Search History". The current page is titled "Choose a Column to Begin Browsing:". Below this, there are several columns of DDC classes, each with a "View Records" link. The classes are:

- 0 Dewey Decimal Classification (DDC) - Document Type
- 1 Philosophy & psychology (41814)
- 2 Religion (18316)
- 3 Social sciences (27940)
- 4 Language (16925)
- 5 Science (253134)
- 6 Technology (314922)
- 7 Arts & recreation (1146711)
- 80 Computer science, knowledge & systems (67818)
- 01 Biographies (156)
- 02 Library & information sciences (11598)
- 03 Encyclopaedias & books of facts (17)
- 05 Magazines, journals & serials (4099)
- 06 Associations, organizations & museums (983)
- 08 Library & information sciences (11578)
- 021 Library relationships (3)
- 025 Library operations (14)
- 026 Libraries for specific subjects (1)
- 027 General libraries (6)

Below the DDC list, there is a section titled "How to browse the DDC" with the following text: "This browsing tool is subdivided into 3 levels. Example: Main class 5 (Natural sciences & mathematics), Division 53 (Physics), Section 539 (Modern physics). Click on an entry to get to the next sub-level. Click on the link 'View Records' to start searching BASE for documents within this main class, division or section. If you search for a main class, the divisions and sections are automatically searched as well, if you search for a division the sections are automatically searched as well."

- Automatische Klassifikation nach der Dewey Dezimalklassifikation im BASE-Kontext

- 1 Motivation
- 2 Wie funktioniert automatische Sacherschließung?**
- 3 Praxisbeispiele UB Bielefeld
- 4 Zusammenfassung

- Die meisten Ansätze der automatisierten Sacherschließung sind heutzutage mit Methoden des maschinellen Lernens realisiert
- Teilbereich der künstlichen Intelligenz (KI)

## Definition nach Mitchell (1997)

A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

- Die meisten Ansätze der automatisierten Sacherschließung sind heutzutage mit Methoden des maschinellen Lernens realisiert
- Teilbereich der künstlichen Intelligenz (KI)

## Definition nach Mitchell (1997)

A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$ , if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .

# Unüberwachtes Lernen

## Zielkategorien nicht bekannt

The screenshot shows the Google News interface in a browser window. The address bar displays 'http://news.google.com/'. The page is in German, with the 'Deutschland-Ausgabe' (Germany Edition) selected. The main headline is '5-Prozent-Hürde bei Europawahl verfassungswidrig' (5% threshold in European election unconstitutional), dated 'Berliner Morgenpost - vor 13 Minuten'. The article text states: 'Das Bundesverfassungsgericht sieht in der Fünf-Prozent-Hürde bei Europawahlen eine Verletzung der Chancengleichheit der Parteien. Der Zweite Senat urteilte am Mittwoch über die Beschwerden des Staatsrechtsprofessors Hans Herbert von Arnim und zweier ...'. Below the article, there are navigation options for 'West Online', 'AFP', 'Handelsblätt', and 'EurActiv.de'. Another headline below is 'Berlusconi Rücktritt: Italien muss sich frei machen' (Berlusconi resignation: Italy must free itself), dated 'STERN.DE - vor 8 Minuten'. The article text says: 'Silvio Berlusconi Rücktritt war überfällig. Viel zu lange hat er Italien und dann auch Europa zu Geiseln seiner Eitelkeit, seines Geldes und seines Unterleibs gemacht. Aber es wäre ein Fehler zu glauben, damit wären auch nur die wichtigsten Probleme ...'. The right sidebar features a video player with a 'HOLLYWOOD' and 'ORGANIC FOOD' logo, and a section titled 'Neueste Nachrichten' (Latest News) with items like 'EKD verbietet auch künftig Streiks' (EKD also bans strikes from now on) and 'Ahmadinejad will Atomprogramm niemals aufgeben' (Ahmadinejad will not give up nuclear program).

- Google News: thematische Gruppierung von Nachrichten

# Unüberwachtes Lernen

## Zielkategorien nicht bekannt

The screenshot shows the Google News interface. The main article is titled "5-Prozent-Hürde bei Europawahl verfassungswidrig" (5% threshold in European election unconstitutional). The article text states: "Das Bundesverfassungsgericht sieht in der Fünf-Prozent-Hürde bei Europawahlen eine Verletzung der Chancengleichheit der Parteien. Der Zweite Senat urteilte am Mittwoch über die Beschwerden des Staatsrechtsprofessors Hans Herbert von Arnim und zweier ...". Below the article, there are several smaller news items, including "Berlusconi Rücktritt: Italien muss sich frei machen" and "EKD verbietet auch künftig Streiks". The interface includes a search bar, navigation tabs, and a sidebar with categories like "Schlagzeilen", "Wirtschaft", and "Sport".

- Google News: thematische Gruppierung von Nachrichten

# Unüberwachtes Lernen

## Zielkategorien nicht bekannt

DLF OAI Portal Search Results

http://quod.lib.umich.edu/cgi/b/bib/bib-idx?c=imis;type=browse;frame=topic;sa=1;sc=8

DLF DIGITAL LIBRARY FEDERATION

Home Search Browse About Contributors Help

You found **155015** records from **Art and Design**.  
Try a **search** to retrieve a different set of records.

Browse by Topic **Browse by Data Contributor**

Arts & Humanities (1014093)

- American and Canadian Studies (183681)
- Anthropology and Archaeology (97399)
- Architecture (300664)
- Art History (86550)
- Art and Design (155015)**
- East Asian Languages and Cultures (34461)
- English Language and Literature (125712)
- History (General) (383174)
- Humanities (General) (10671)
- Linguistics (10398)
- Music and Dance (54862)
- Philosophy (41549)
- Religious Studies (7217)
- Russian and East European Studies (36714)

Jump to Records: 1 | 11 | 21 ... 15501 ... 31001 ... 46501 ... 62001 ... 77501 ... 93001 ... 108501 ... 124011 ... 139511 ... 155011

Next 10 Records ... 139511 ... 155011

**Record 1 of 155015**  
add to bookbag

Title	Stray Pieces of Early Christian Writing
Author/Creator	Sherman E. Johnson
Publisher	University of Chicago Press
Year	2003-01-13T16:22:49Z
Resource Type	journal article
Resource Format	application/pdf
Language	English
Source	Journal of Near Eastern Studies, Volume 5, Page 40
Note	10.1086/370769
Subject	epstein-barr virus; weaving; historical collection; vologe
Subject	Art and Design; Internal Medicine and Specialties; Electrical Engineering; History (General)
Subject	Science; Arts & Humanities; Engineering; Social Sciences; Health Sciences
URL	http://www.journals.uchicago.edu/cgi-bin/resolve?id=doi:10.1086/370769
Data Contributor	The University of Chicago Press Journals Division

**Record 2 of 155015**  
add to bookbag

Title	The Artist of the Egyptian Old Kingdom
Author/Creator	John A. Wilson
Publisher	University of Chicago Press

- UL Michigan: Thematische Gruppierung von Metadaten + anschließendes Labeling

# Überwachtes Lernen

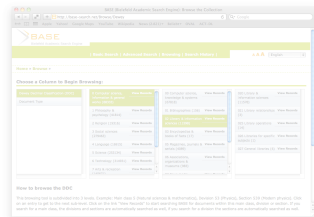
Zielkategorien bekannt



Spam-Filter



Klassifikation nach  
DNB-Sachgruppen



Klassifikation nach DDC



# Überwachtes Lernen

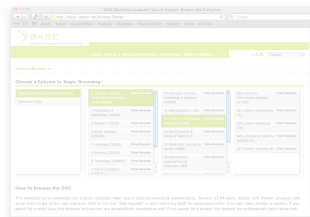
Zielkategorien bekannt



Spam-Filter

DEUTSCHE  
NATIONAL  
BIBLIOTHEK

Klassifikation nach  
DNB-Sachgruppen



Klassifikation nach DDC

# Überwachtes Lernen

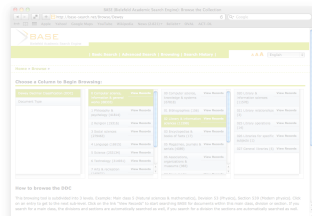
Zielkategorien bekannt



Spam-Filter



Klassifikation nach  
DNB-Sachgruppen



Klassifikation nach DDC

# Überwachtes Lernen

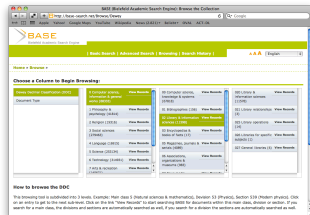
Zielkategorien bekannt



Spam-Filter



Klassifikation nach  
DNB-Sachgruppen

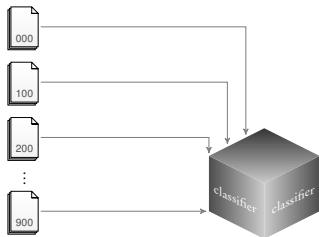


Klassifikation nach DDC

# Klassifikation nach DDC

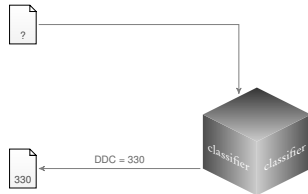
## Lernphase

Training examples



## Applikationsphase

Unclassified document



**Task T:** Klassifiziere beliebige Textdokumente nach der DDC

**Experience E:** DDC-klassifizierte Beispieldokumente

**Performance P:** z.B. Anteil der korrekt klassifizierten Dokumente

- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
  - Metadaten und Volltexte
  - Ermittlung von DDC-Nummern über Konkordanzen
  - ~100.000 DDC-Kategorisierte Trainingsdokumente

- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
- Metadaten und Volltexte
- Ermittlung von DDC-Nummern über Konkordanzen
- ~100.000 DDC-Kategorisierte Trainingsdokumente

- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
- Metadaten und Volltexte
- Ermittlung von DDC-Nummern über Konkordanzen
- ~100.000 DDC-Kategorisierte Trainingsdokumente

- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
- Metadaten und Volltexte
- Ermittlung von DDC-Nummern über Konkordanzen
- ~100.000 DDC-Kategorisierte Trainingsdokumente



- 1 Motivation
- 2 Wie funktioniert automatische Sacherschließung?
- 3 Praxisbeispiele UB Bielefeld**
- 4 Zusammenfassung

# Metadatenanreicherung in BASE

## OAI-DC-Metadaten

```
<record>
...
<metadata>
  <oai_dc:dc xmlns:xsi="...">
    <dc:title>
      Bielefeld Academic Search Engine (BASE): an end-user oriented
      institutional repository search service
    </dc:title>
    <dc:creator>Pieper, Dirk</dc:creator>
    <dc:creator>Summann, Friedrich</dc:creator>
    <dc:subject>LS. Search engines.</dc:subject>
    <dc:subject>HS. Repositories.</dc:subject>
    <dc:description>
      Purpose: This paper describes the activities of Bielefeld University
      Library in establishing OAI based repository servers and in using OAI
      resources for end-user-oriented search services like BASE (Bielefeld
      Academic Search Engine). Design/methodology/approach: BASE uses the
      search engine technology Fast Data Search. Findings: BASE is able to
      integrate external functions of Google Scholar. The search engine
      technology can replace or amend the search functions of a given
      repository software. BASE can also be embedded in external repository
      environments. Originality/value: The paper provides an overview over
      the functionalities of BASE and gives insight into the challenges that
      have to be faced when harvesting and integrating resources from multiple
      OAI servers.
    </dc:description>
    <dc:publisher>Emerald</dc:publisher>
    <dc:date>2006</dc:date>
    <dc:type>Journal Article (Print/Paginated)</dc:type>
    <dc:type>PeerReviewed</dc:type>
    <dc:format>application/pdf</dc:format>
    <dc:relation>
      http://conference.ub.uni-bielefeld.de/2006/proceedings/pieper_summann_final_web.pdf
    </dc:relation>
    <dc:identifier>http://eprints.rclis.org/9160</dc:identifier>
    <dc:language>en</dc:language>
  </oai_dc:dc>
</metadata>
</record>
```

# Metadatenanreicherung in BASE

## OAI-DC-Metadaten

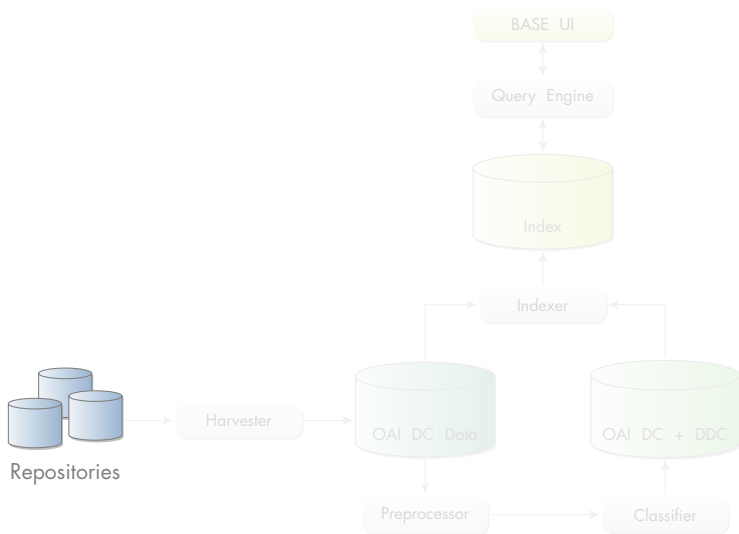
```
<record>
...
<metadata>
  <oai_dc:dc xmlns:xsi="...">
    <dc:title>
      Bielefeld Academic Search Engine (BASE): an end-user oriented
      institutional repository search service
    </dc:title>
    <dc:creator>Pieper, Dirk</dc:creator>
    <dc:creator>Summann, Friedrich</dc:creator>
    <dc:subject>LS. Search engines.</dc:subject>
    <dc:subject>HS. Repositories.</dc:subject>
    <dc:description>
      Purpose: This paper describes the activities of Bielefeld University
      Library in establishing OAI based repository servers and in using OAI
      resources for end-user-oriented search services like BASE (Bielefeld
      Academic Search Engine). Design/methodology/approach: BASE uses the
      search engine technology Fast Data Search. Findings: BASE is able to
      integrate external functions of Google Scholar. The search engine
      technology can replace or amend the search functions of a given
      repository software. BASE can also be embedded in external repository
      environments. Originality/value: The paper provides an overview over
      the functionalities of BASE and gives insight into the challenges that
      have to be faced when harvesting and integrating resources from multiple
      OAI servers.
    </dc:description>
    <dc:publisher>Emerald</dc:publisher>
    <dc:date>2006</dc:date>
    <dc:type>Journal Article (Print/Paginated)</dc:type>
    <dc:type>PeerReviewed</dc:type>
    <dc:format>application/pdf</dc:format>
    <dc:relation>
      http://conference.ub.uni-bielefeld.de/2006/proceedings/pieper_summann_final_web.pdf
    </dc:relation>
    <dc:identifier>http://eprints.rclis.org/9160</dc:identifier>
    <dc:language>en</dc:language>
  </oai_dc:dc>
</metadata>
</record>
```

# Metadatenanreicherung in BASE

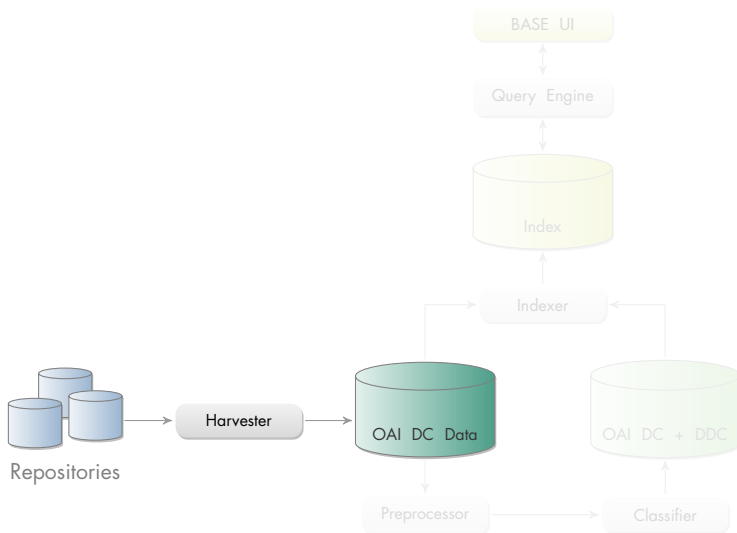
## OAI-DC-Metadaten

```
<metadata>
  <oai_dc:dc xmlns:xsi="...">
    <dc:title>
      Bielefeld Academic Search Engine (BASE): an end-user oriented
      institutional repository search service
    </dc:title>
    <dc:creator>Pieper, Dirk</dc:creator>
    <dc:creator>Summann, Friedrich</dc:creator>
    <dc:subject>LS. Search engines.</dc:subject>
    <dc:subject>HS. Repositories.</dc:subject>
    <dc:description>
      Purpose: This paper describes the activities of Bielefeld
      Library in establishing OAI based repository servers and
      resources for end-user-oriented search services like BASE
      Academic Search Engine). Design/methodology/approach: BASE
      search engine technology Fast Data Search. Findings: BASE
      integrate external functions of Google Scholar. The search
      technology can replace or amend the search functions of a
      repository software. BASE can also be embedded in external
      environments. Originality/value: The paper provides an overview
      the functionalities of BASE and gives insight into the challenges
      have to be faced when harvesting and integrating resources from
      OAI servers.
    </dc:description>
  </oai_dc:dc>
</metadata>
```

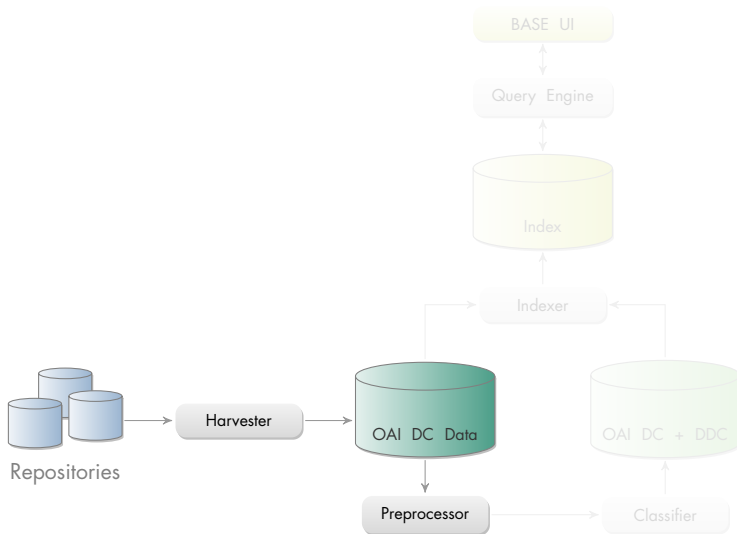
# Metadatenanreicherung in BASE



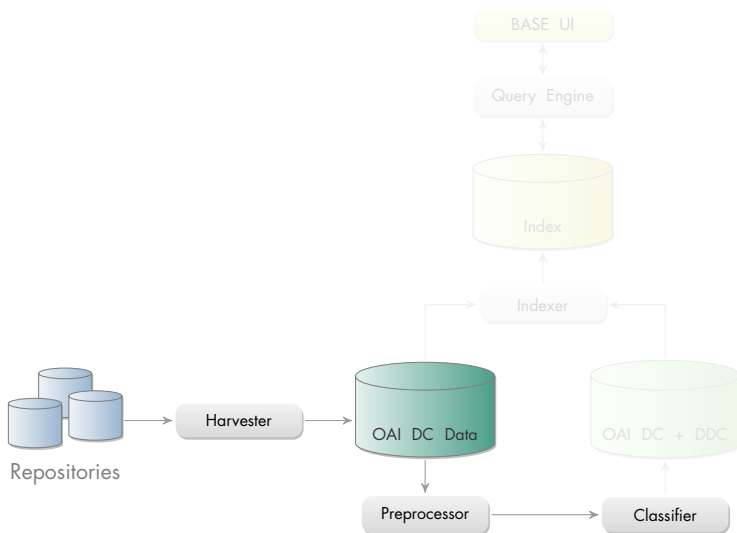
# Metadatenanreicherung in BASE



# Metadatenanreicherung in BASE

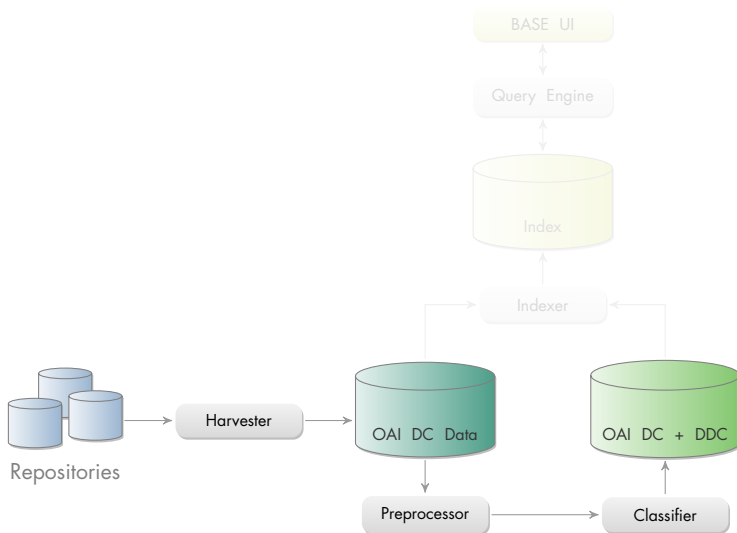


# Metadatenanreicherung in BASE

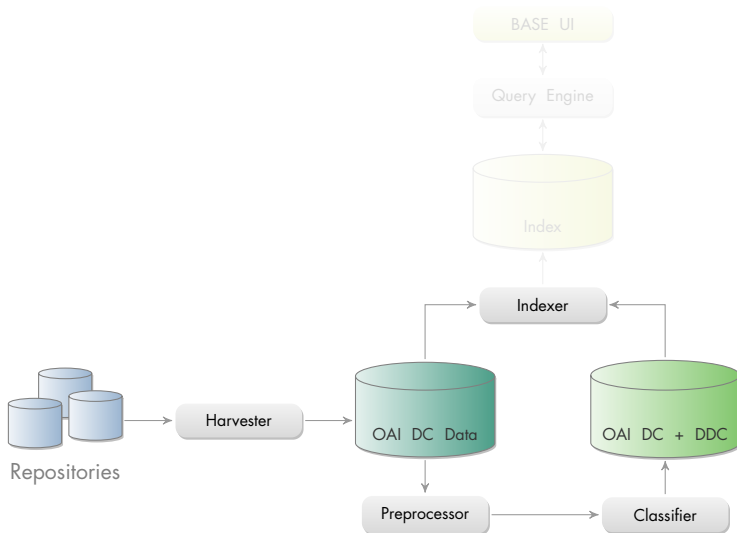




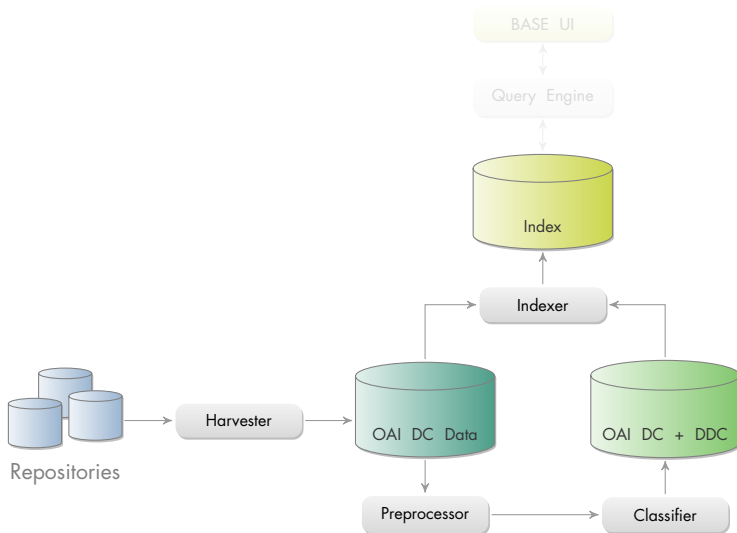
# Metadatenanreicherung in BASE



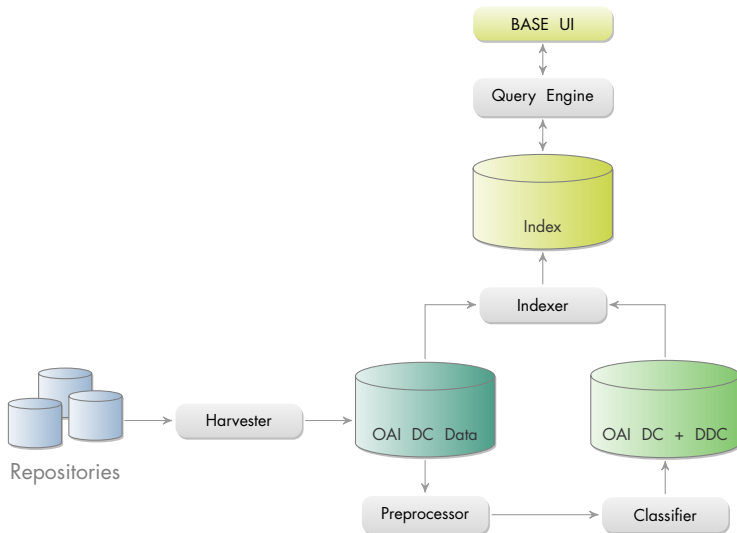
# Metadatenanreicherung in BASE



# Metadatenanreicherung in BASE



# Metadatenanreicherung in BASE



# BASE Browsing Interface

Entwickelt von Renata Mitrenga

The screenshot shows a web browser window with the URL <http://base-search.net/Browse/Dewey>. The page title is "BASE (Bielefeld Academic Search Engine): Browse the Collection". The navigation menu includes "Basic Search", "Advanced Search", "Browsing", and "Search History". The language is set to "English".

The main content area is titled "Choose a Column to Begin Browsing:" and displays a grid of Dewey Decimal Classification (DDC) categories. Each category is presented in a box with a title, a description, and a "View Records" link. The categories are:

- 0 Computer science, information & general works (88333)
- 1 Philosophy & psychology (41814)
- 2 Religion (19316)
- 3 Social sciences (279463)
- 4 Language (16915)
- 5 Science (253134)
- 6 Technology (314931)
- 7 Arts & recreation (148471)
- 00 Computer science, knowledge & systems (67818)
- 01 Bibliographies (156)
- 02 Library & information sciences (11598)
- 03 Encyclopedias & books of facts (17)
- 05 Magazines, journals & serials (4089)
- 06 Associations, organizations & museums (983)
- 020 Library & information sciences (11578)
- 021 Library relationships (3)
- 025 Library operations (14)
- 026 Libraries for specific subjects (1)
- 027 General libraries (6)

Below the grid, there is a section titled "How to browse the DDC" with the following text:

This browsing tool is subdivided into 3 levels. Example: Main class 5 (Natural sciences & mathematics), Devision 53 (Physics), Section 539 (Modern physics). Click on an entry to get to the next sub-level. Click on the link "View Records" to start searching BASE for documents within this main class, division or section. If you search for a main class, the divisions and sections are automatically searched as well, if you search for a division the sections are automatically searched as well.

# Feedback-Formular

Entwickelt von Renata Mitrenga

[schließen](#)

### DDC Korrekturvorschlag

---

Bitte wählen Sie die korrekte DDC-Klasse für dieses Dokument in drei Schritten aus:

Hinweis: Ihr Vorschlag hilft uns bei der Verbesserung der automatischen Klassifikation. Eine direkte Übernahme in die BASE-Ergebnisse findet nicht statt.

- Korrekturvorschläge von Nutzern für Kategorisierung
- Korrigierte Dokumente werden für das Training verwendet

# Feedback-Formular

Entwickelt von Renata Mitrenga

[schließen](#)

### DDC Korrekturvorschlag

---

Bitte wählen Sie die korrekte DDC-Klasse für dieses Dokument in drei Schritten aus:

▾

▾

▾

Hinweis: Ihr Vorschlag hilft uns bei der Verbesserung der automatischen Klassifikation. Eine direkte Übernahme in die BASE-Ergebnisse findet nicht statt.

- Korrekturvorschläge von Nutzern für Kategorisierung
- Korrigierte Dokumente werden für das Training verwendet

Aktuelle Zahlen aus BASE:

**Intellektuell klassifizierte Dokumente:** 429.322

Automatisch klassifizierte Dokumente: 846.143

DDC-klassifiziert total: 1.275.465

Steigerung durch Automatisierung: 197,09 %



# Ergebnisse

## Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

**Intellektuell klassifizierte Dokumente:** 429.322

**Automatisch klassifizierte Dokumente:** 846.143

DDC-klassifiziert total: 1.275.465

Steigerung durch Automatisierung: 197,09%

# Ergebnisse

## Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

**Intellektuell klassifizierte Dokumente:** 429.322

**Automatisch klassifizierte Dokumente:** 846.143

**DDC-klassifiziert total:** 1.275.465

**Steigerung durch Automatisierung:** 197,09%

# Ergebnisse

## Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

**Intellektuell klassifizierte Dokumente:** 429.322

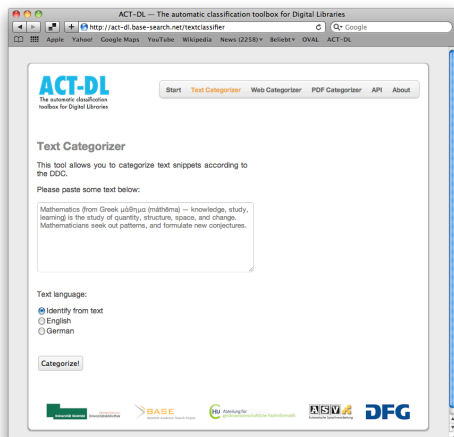
**Automatisch klassifizierte Dokumente:** 846.143

**DDC-klassifiziert total:** 1.275.465

**Steigerung durch Automatisierung:** 197,09 %

# ACT-DL

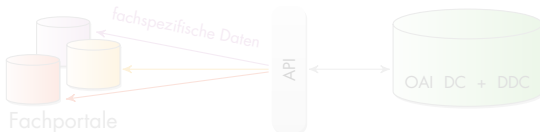
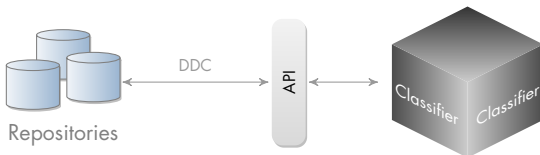
## The Automatic Classification Toolbox for Digital Libraries



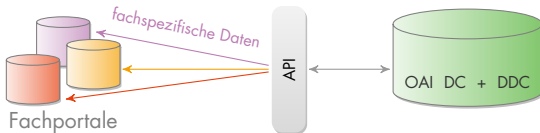
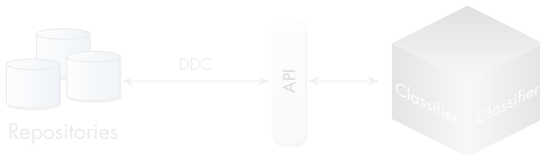
- Text Categorizer
- Web Categorizer
- PDF Categorizer

▶ Seite besuchen

# Nachnutzung des Klassifikators



# Nachnutzung klassifizierter Metadaten



# Nachnutzung: Klassifikator

## API

```
<results language="en">
  <result level="1">
    <DDC number="5" heading="Science" confidence="0.642842433048"/>
    <DDC number="3" heading="Social sciences" confidence="0.153678304171"/>
    <DDC number="6" heading="Technology" confidence="0.113899075409"/>
    <DDC number="7" heading="Arts & recreation" confidence="0.0509034274529"/>
    <DDC number="0" heading="Computer science ..." confidence="..."/>
    <DDC number="9" heading="History & geography" confidence="0.00924913669985"/>
    <DDC number="2" heading="Religion" confidence="0.00839946320403"/>
    <DDC number="8" heading="Literature" confidence="0.00567627025569"/>
    <DDC number="4" heading="Language" confidence="0.00295446462836"/>
    <DDC number="1" heading="Philosophy & psychology" confidence="0.00199558448278"/>
  </result>
  <result level="2">
    <DDC number="51" heading="Mathematics" confidence="0.948098700683"/>
    <DDC number="57" heading="Life sciences; biology" confidence="0.0121094418067"/>
    <DDC number="59" heading="Animals (Zoology)" confidence="0.00915736550968"/>
    <DDC number="53" heading="Physics" confidence="0.00896283232273"/>
    <DDC number="54" heading="Chemistry" confidence="0.00890949865225"/>
    <DDC number="50" heading="Science" confidence="0.00553605223902"/>
    <DDC number="58" heading="Plants (Botany)" confidence="0.00287309981003"/>
    <DDC number="55" heading="Earth sciences & geology" confidence="0.00250678827777"/>
    <DDC number="52" heading="Astronomy" confidence="0.00108935568839"/>
    <DDC number="56" heading="Fossils & prehistoric life" confidence="..."/>
  </result>
  <result level="3">
    <DDC number="510" heading="Mathematics" confidence="0.987321567068"/>
    <DDC number="515" heading="Analysis" confidence="0.0036012222724"/>
    <DDC number="518" heading="Numerical analysis" confidence="0.00244515445432"/>
    <DDC number="512" heading="Algebra" confidence="0.00229963903671"/>
    <DDC number="516" heading="Geometry" confidence="0.00223162097482"/>
    <DDC number="519" heading="Probabilities & applied mathematics" confidence="..."/>
  </result>
</results>
```

# Nachnutzung des Klassifikators in PUB

## Vorschlagsystem für die Metadatenerfassung

Publications at Bielefeld University  
http://bup-dev.ub.uni-bielefeld.de/luur/Record

**"Hierarchical Classification of OAI Metadata Using the DDC Taxonomy" (Book Chapter)**

Work   Abstract + Subject   Fulltext   Message   Show all tabs on one page

**Abstract + Subject**

Abstract + In the area of digital library services, the access to subject-specific metadata of scholarly publications is of utmost interest. One of the most prevalent approaches for metadata exchange is the XML-based Open Archive Initiative (OAI) Protocol for Metadata Harvesting (OAI-PMH). However, due to its loose requirements regarding metadata content there is no strict standard for consistent subject

Language of Abstract English

Keywords Dewey Decimal Classification; Digital Library; OAI-PMH; SVM; Hierarchical Classification

Subject Technology and Engineering

DDC 020 Library & Information sciences Suggest DDC

References

Save Change Type Return Delete Close

Fertig



# Nachnutzung des Klassifikators in PUB

Vorschlagsystem für die Metadatenerfassung

Language of Abstract: English

Dewey Decimal Classification; Digital Library; OAI-PMH; SVM;  
Hierarchical Classification

Technology and Engineering

-- 020 Library & information sciences

Suggest DDC

Change Type

Return

Delete

Close

- Schnittstelle für Fachportale für den fachspezifischen Metadatenaustausch
- Pilotpartner: *EconBiz.de* (Virtuelle Fachbibliothek Wirtschaftswissenschaften, ZBW Kiel)
- Zusammenarbeit mit dem DFG-Projekt *Open Access Fachrepositorien OAFR* (UB Konstanz)

- 1 Motivation
- 2 Wie funktioniert automatische Sacherschließung?
- 3 Praxisbeispiele UB Bielefeld
- 4 **Zusammenfassung**

- Automatische Sacherschließung mit maschinellen Lernverfahren gewinnt an Bedeutung im Bibliotheksbereich
- UB hat mit BASE ideale Voraussetzungen für Entwicklungen in diesem Bereich:
  - großer Korpus an Trainings- und Testdokumenten
  - Showcase für entwickelte Applikationen
- Vielfältige Möglichkeiten der Nachnutzung:
  - Repositories (PUB): Unterstützung bei der Metadatenerfassung
  - virtuelle Fachportale: Belieferung mit fachspezifischen Metadaten
  - FP7-Projekt OpenAIREplus: Workpackage zur automatischen Klassifikation

- Automatische Sacherschließung mit maschinellen Lernverfahren gewinnt an Bedeutung im Bibliotheksbereich
- UB hat mit BASE ideale Voraussetzungen für Entwicklungen in diesem Bereich:
  - großer Korpus an Trainings- und Testdokumenten
  - Showcase für entwickelte Applikationen
- Vielfältige Möglichkeiten der Nachnutzung:
  - Repositories (PUB): Unterstützung bei der Metadatenerfassung
  - virtuelle Fachportale: Belieferung mit fachspezifischen Metadaten
  - FP7-Projekt OpenAIREplus: Workpackage zur automatischen Klassifikation

- Automatische Sacherschließung mit maschinellen Lernverfahren gewinnt an Bedeutung im Bibliotheksbereich
- UB hat mit BASE ideale Voraussetzungen für Entwicklungen in diesem Bereich:
  - großer Korpus an Trainings- und Testdokumenten
  - Showcase für entwickelte Applikationen
- Vielfältige Möglichkeiten der Nachnutzung:
  - Repositories (PUB): Unterstützung bei der Metadatenerfassung
  - virtuelle Fachportale: Belieferung mit fachspezifischen Metadaten
  - FP7-Projekt OpenAIREplus: Workpackage zur automatischen Klassifikation

Vielen Dank für die Aufmerksamkeit!

Universität Bielefeld | Universitätsbibliothek  
Universitätsstr. 25  
D-33615 Bielefeld  
☎ +49 521 106-2546  
✉ [Mathias.Loesch@uni-bielefeld.de](mailto:Mathias.Loesch@uni-bielefeld.de)

- Hagedorn, K., S. Chapman, and D. Newman (2007). Enhancing search and browse using automated clustering of subject metadata. *D-Lib Magazine* 13(7/8).
- Mehler, A. and U. Waltinger (2009). Enhancing document modeling by means of open topic models: Crossing the frontier of classification schemes in digital libraries by example of the DDC. *Library Hi Tech* 27(4), 520–539.
- Mitchell, T. M. (1997). *Machine learning*. Mcgraw-Hill Higher Education.
- Schöning-Walter, C. (2010). PETRUS – Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek. *Dialog mit Bibliotheken* 1, 15–19.
- Waltinger, U., A. Mehler, M. Lösch, and W. Horstmann (2011). Hierarchical classification of OAI metadata using the DDC taxonomy. In R. Bernardi, S. Chambers, B. Gottfried, F. Segond, and I. Zaihrayeu (Eds.), *Advanced Language Technologies for Digital Libraries*, Volume 6699 of *Lecture Notes in Computer Science*, pp. 29–40. Springer Berlin / Heidelberg.