Automatische Klassifikation in der Praxis

Mathias Lösch

Universitätsbibliothek Bielefeld

Mathias.Loesch@uni-bielefeld.de

Kolloquium Wissensinfrastruktur WS 2011/12
11/11/11

Agenda

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

1 **Motivation**

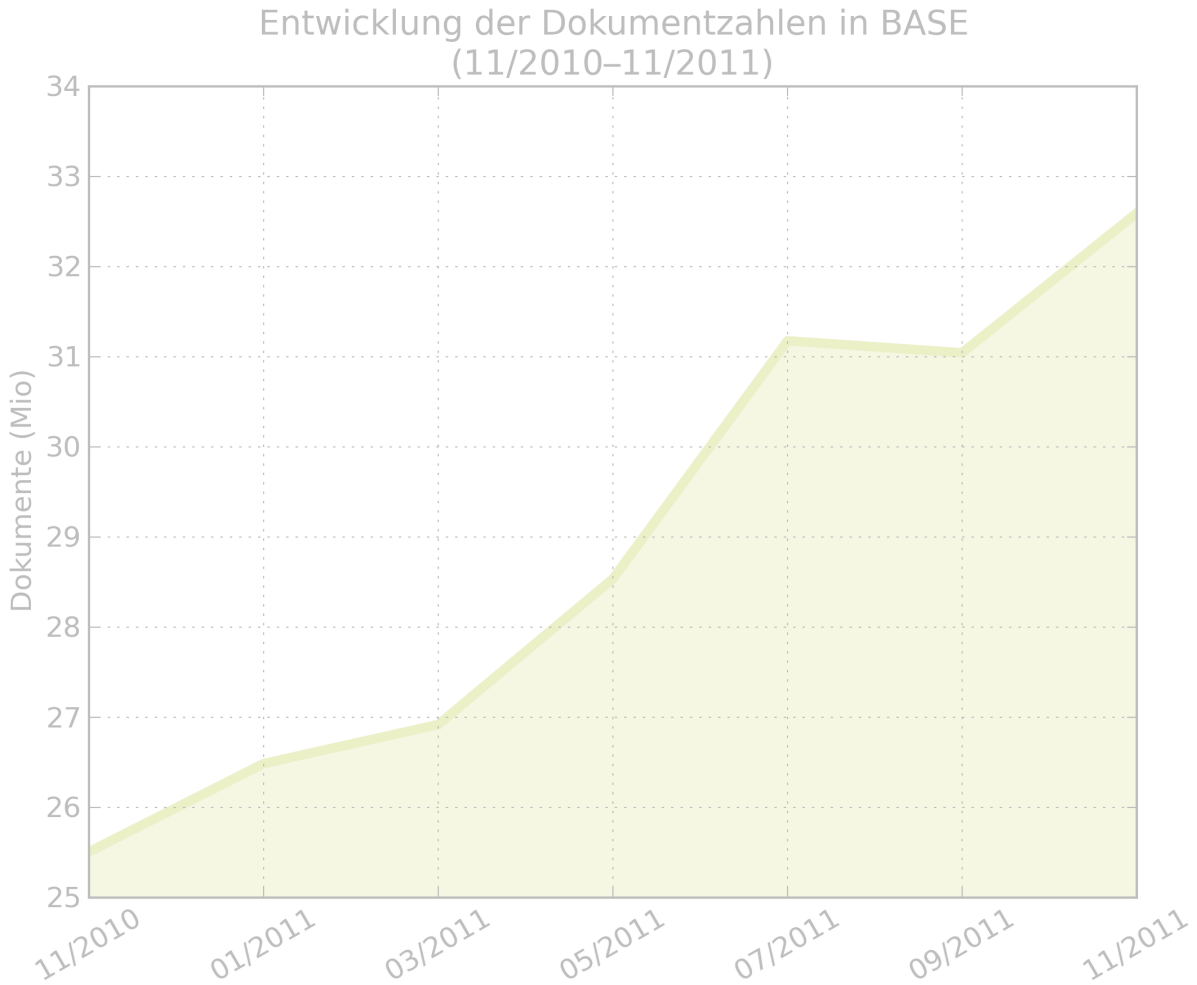
2 Wie funktioniert automatische Sacherschließung?

3 Praxisbeispiele UB Bielefeld

4 Zusammenfassung

Motivation: Information Overload

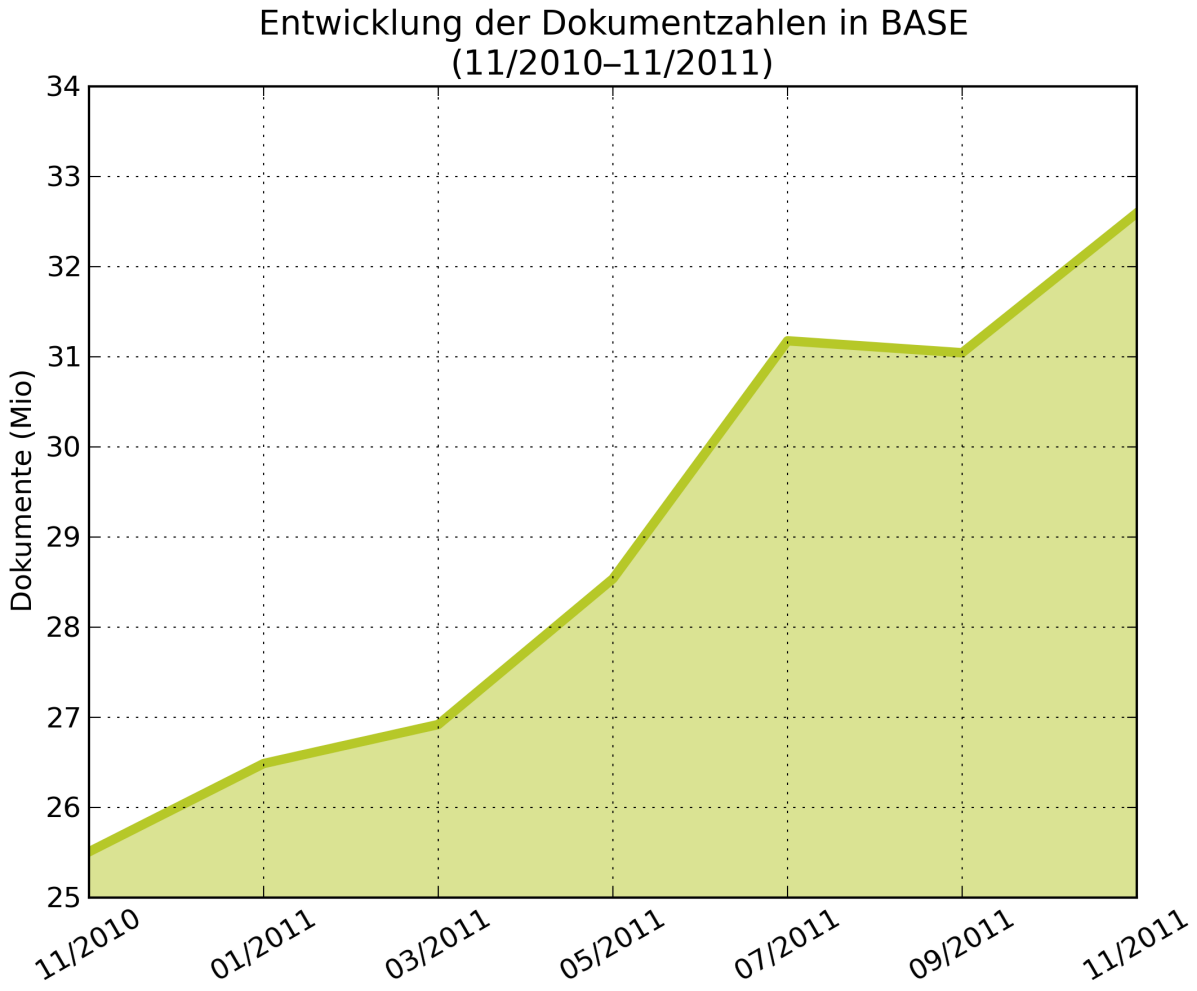
- Starke Zunahme elektronischer Publikationen
- Beispiel BASE: Zunahme nachgewiesener Dokumente seit November 2010 um ~7 Mio.
- ~19.000 Dokumente pro Tag



Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

Motivation: Information Overload

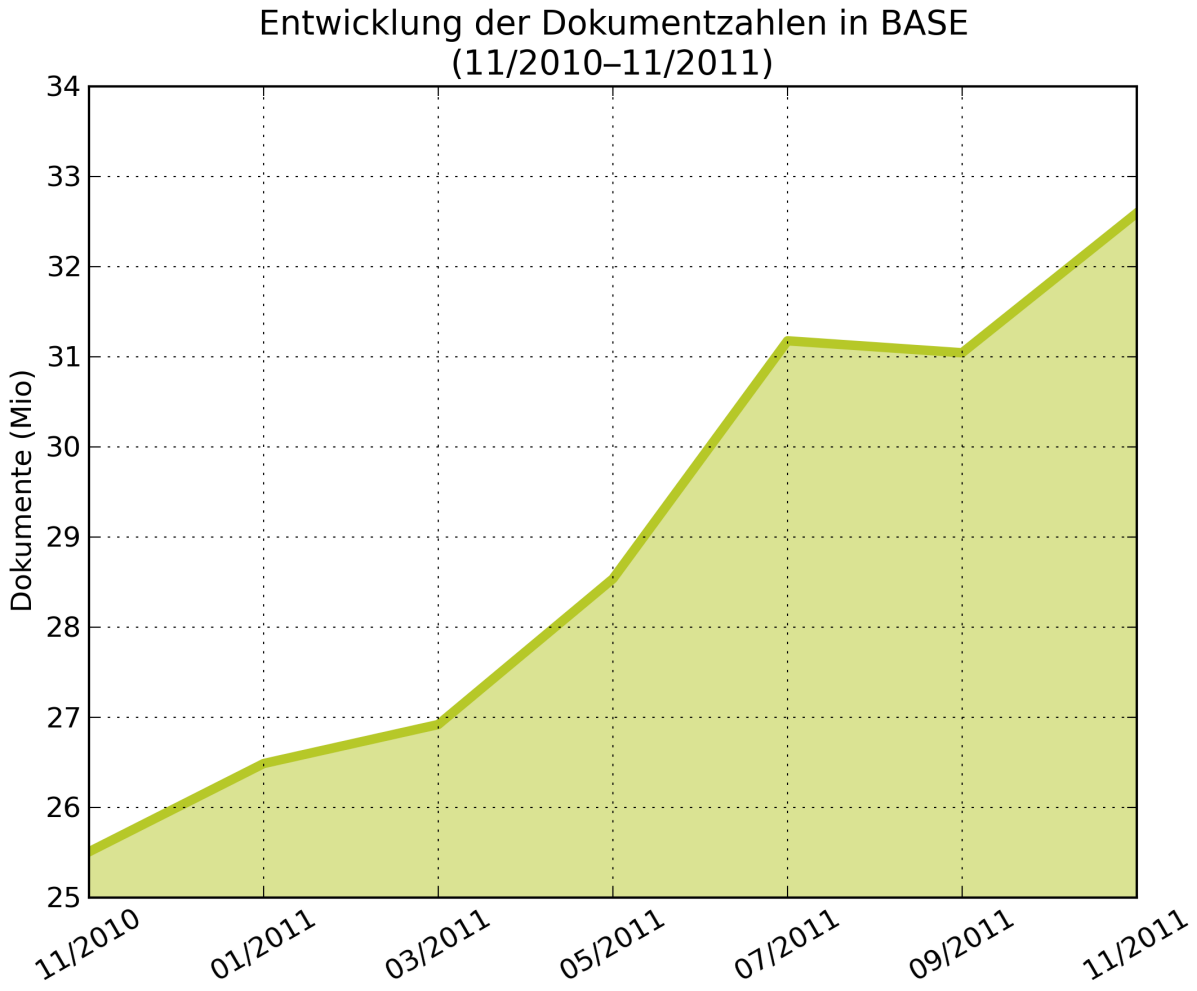
- Starke Zunahme elektronischer Publikationen
- Beispiel BASE: Zunahme nachgewiesener Dokumente seit November 2010 um ~7 Mio.
- ~19.000 Dokumente pro Tag



Universität Bielefeld
Information.plus! Universitätsbibliothek
BASE
Bielefeld Academic Search Engine

Motivation: Information Overload

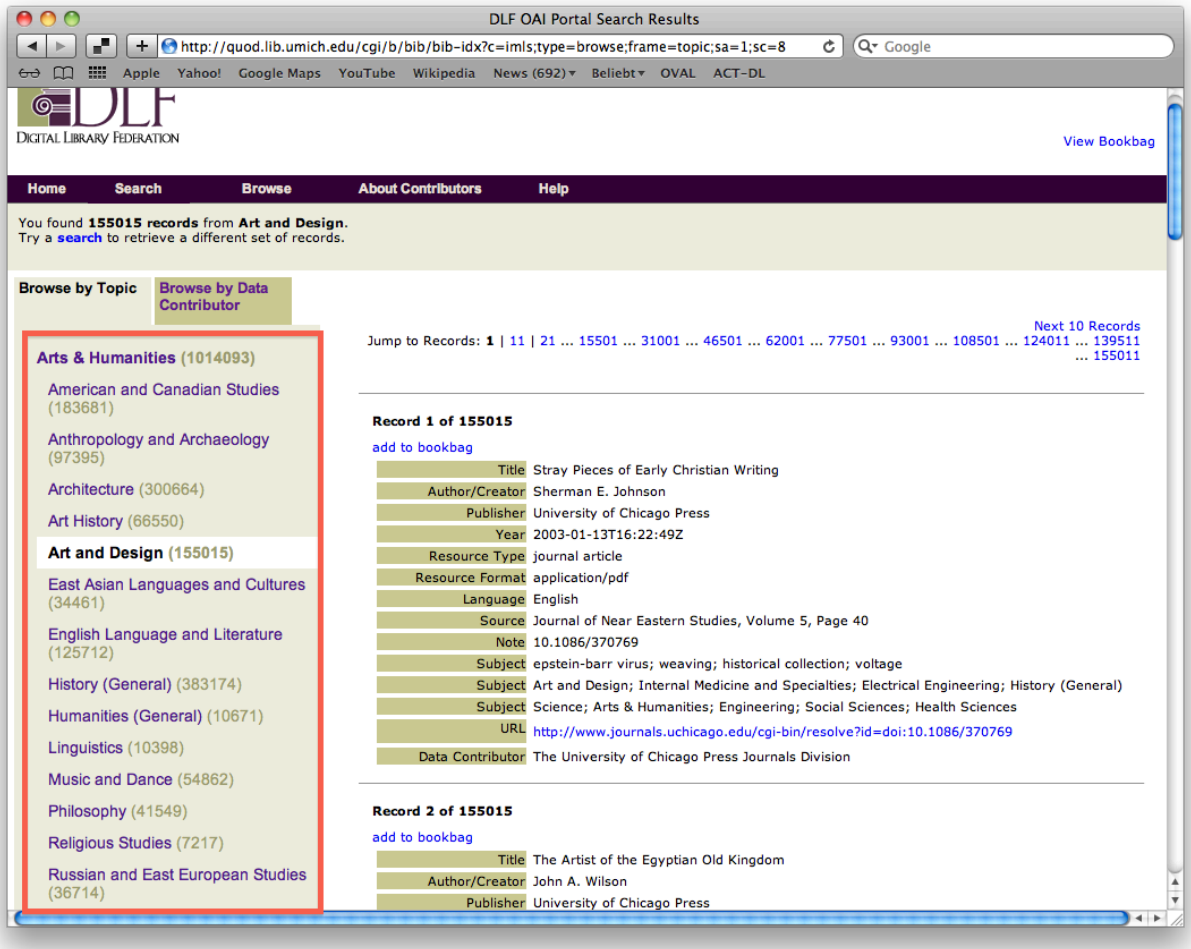
- Starke Zunahme elektronischer Publikationen
- Beispiel BASE: Zunahme nachgewiesener Dokumente seit November 2010 um ~7 Mio.
- ~19.000 Dokumente pro Tag



Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

Automatische Sacherschließung im Bibliothekskontext

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



- Automatische Zuordnung von Datensätzen zu Themen (Topics)

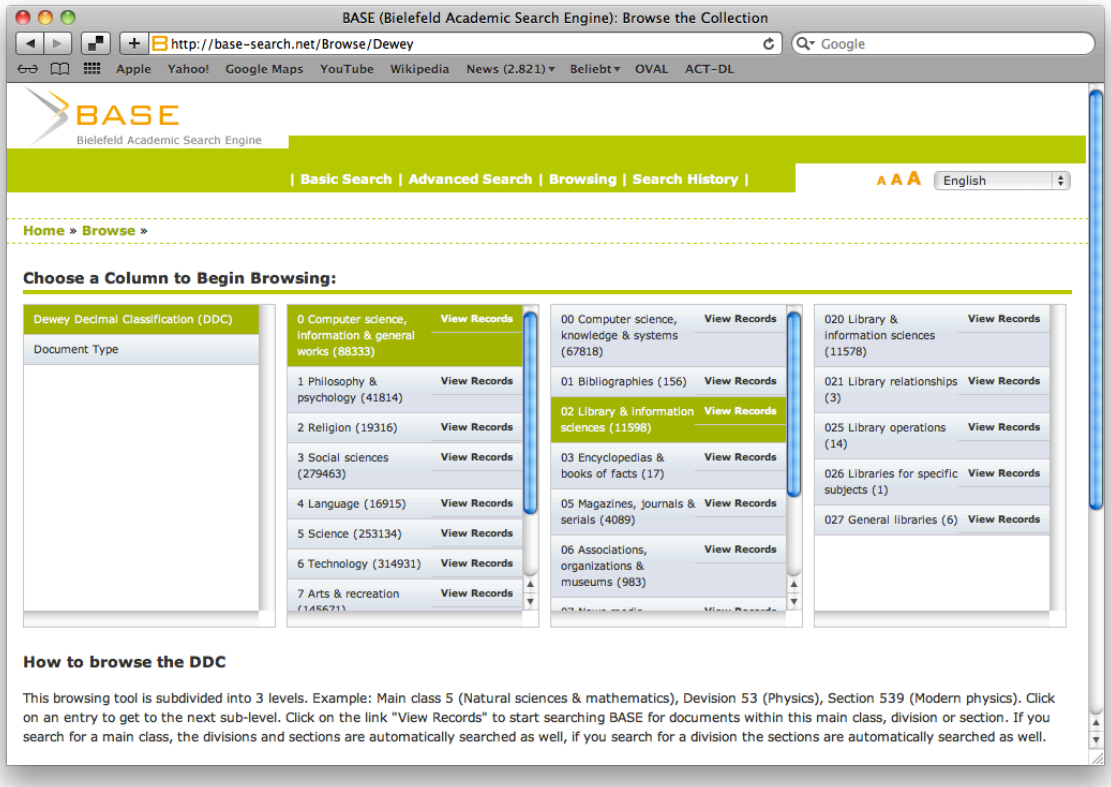
Automatische Sacherschließung im Bibliothekskontext

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«



- Automatische Klassifikation nach der Dewey Dezimalklassifikation im BASE-Kontext

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

1 Motivation

2 Wie funktioniert automatische Sacherschließung?

3 Praxisbeispiele UB Bielefeld

4 Zusammenfassung

Maschinelles Lernen

- 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*.

Maschinelles Lernen

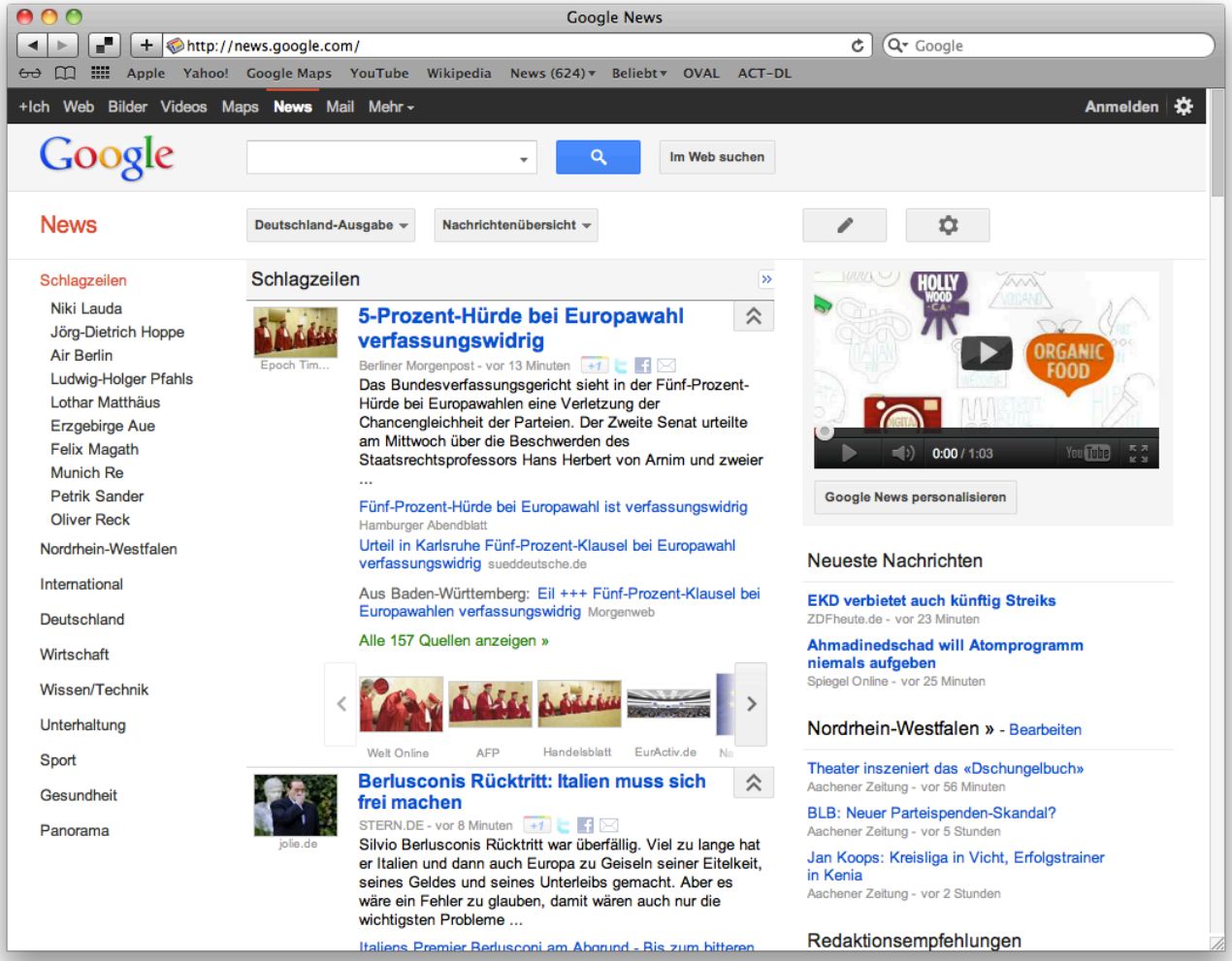
- Die meisten Ansätze der automatisierten Sacherschließung sind heutzutage mit Methoden des maschinellen Lernens realisiert
- Teilbereich der künstlichen Intelligenz (KI)

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

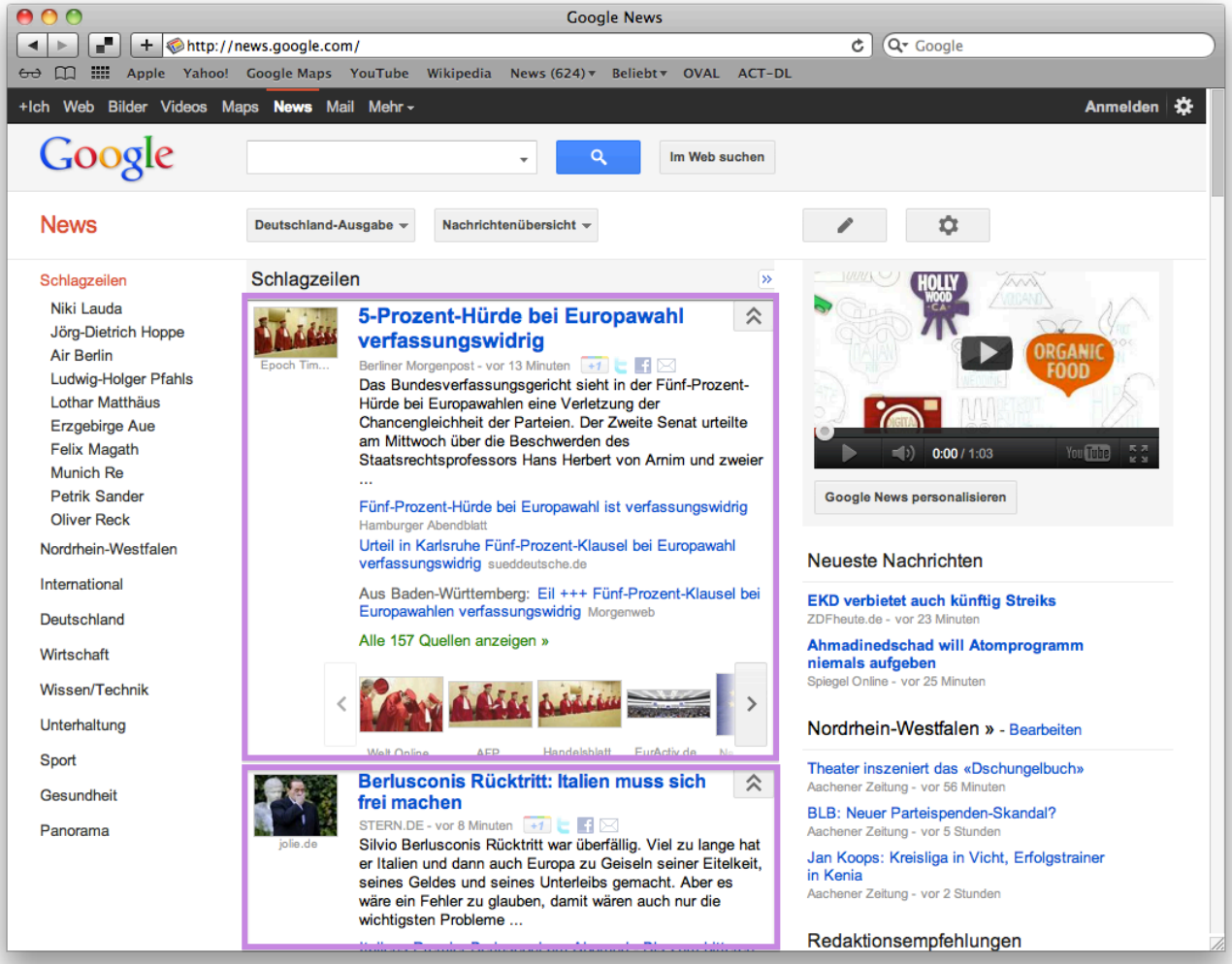
Zielkategorien nicht bekannt



- Google News: thematische Gruppierung von Nachrichten

Unüberwachtes Lernen

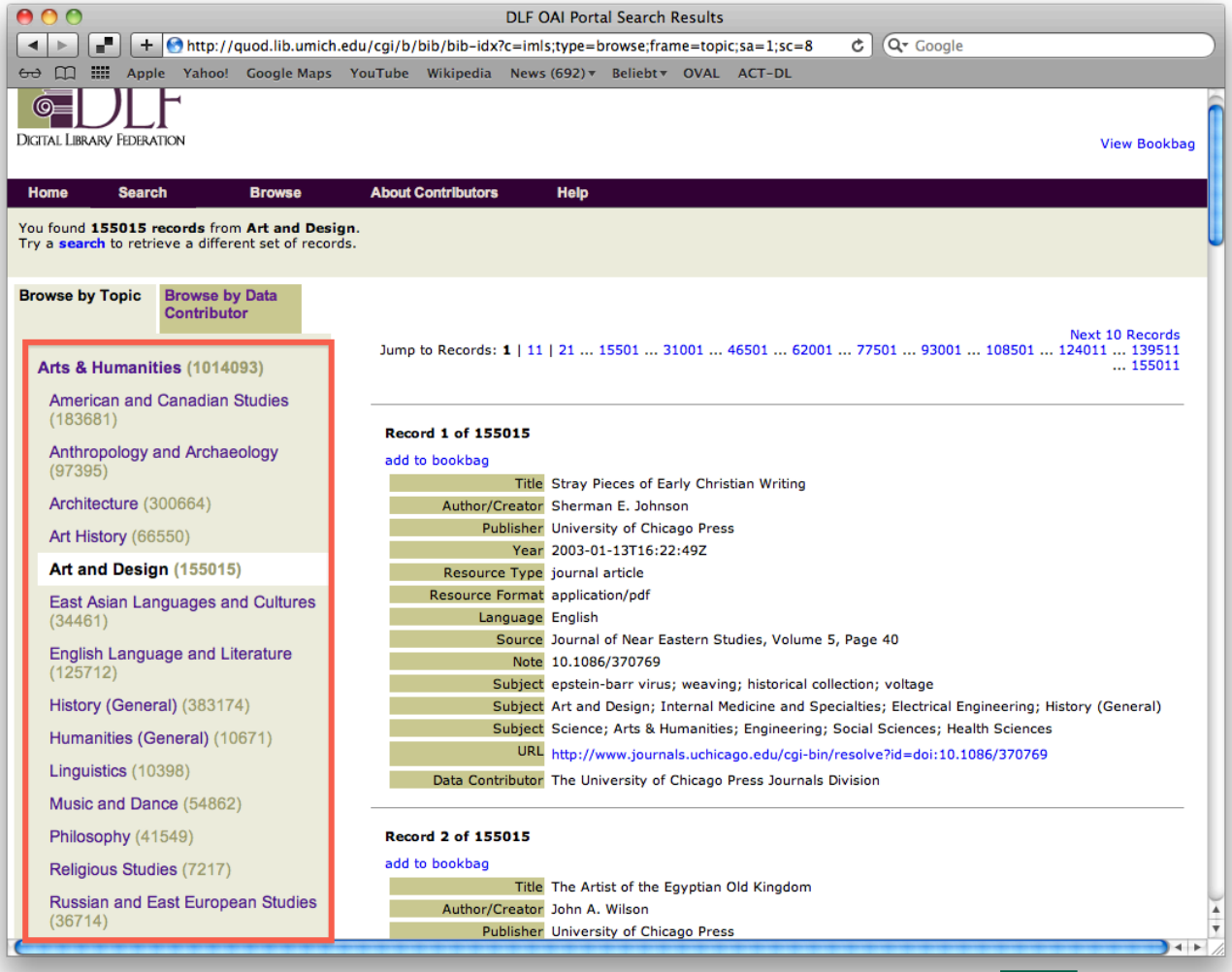
Zielkategorien nicht bekannt



- Google News: thematische Gruppierung von Nachrichten

Unüberwachtes Lernen

Zielkategorien nicht bekannt



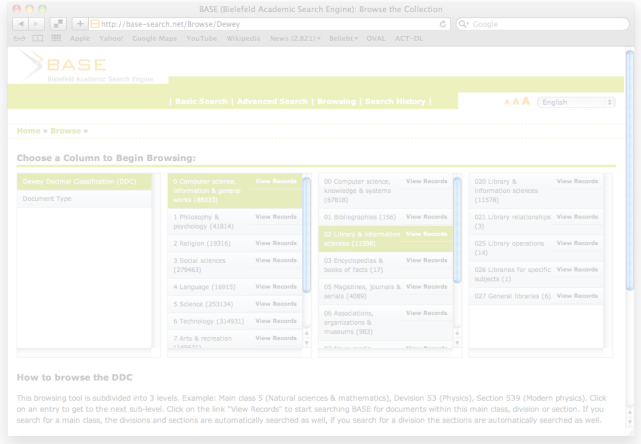
- UL Michigan: Thematische Gruppierung von Metadaten + anschließendes Labeling

Überwachtes Lernen

Zielkategorien bekannt

Spam-Filter

Klassifikation nach
DNB-Sachgruppen



Klassifikation nach DDC

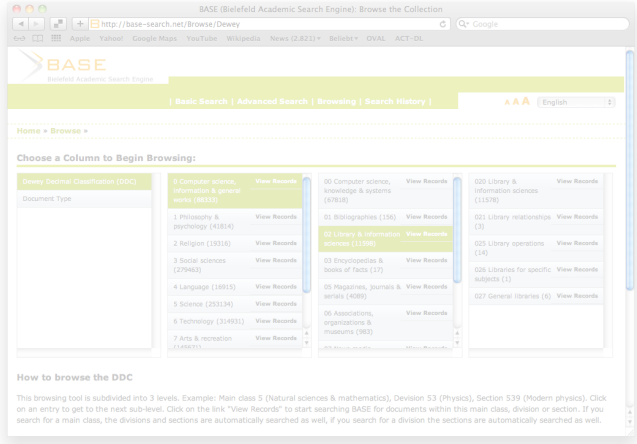
Überwachtes Lernen

Zielkategorien bekannt

Spam-Filter

DEUTSCHE NATIONAL BIBLIOTHEK

Klassifikation nach DNB-Sachgruppen



Klassifikation nach DDC

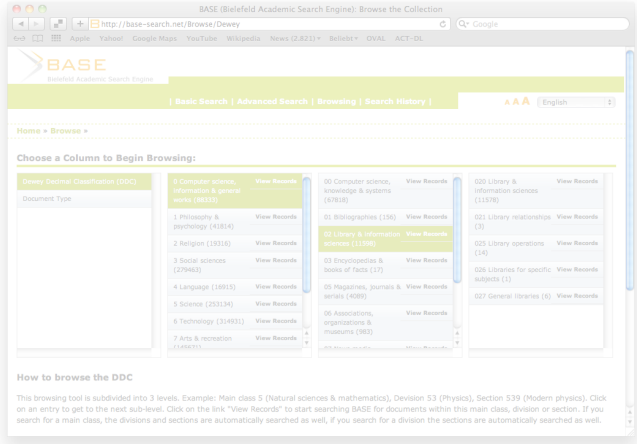
Überwachtes Lernen

Zielkategorien bekannt

Spam-Filter

DEUTSCHE NATIONAL BIBLIOTHEK

Klassifikation nach DNB-Sachgruppen



Klassifikation nach DDC

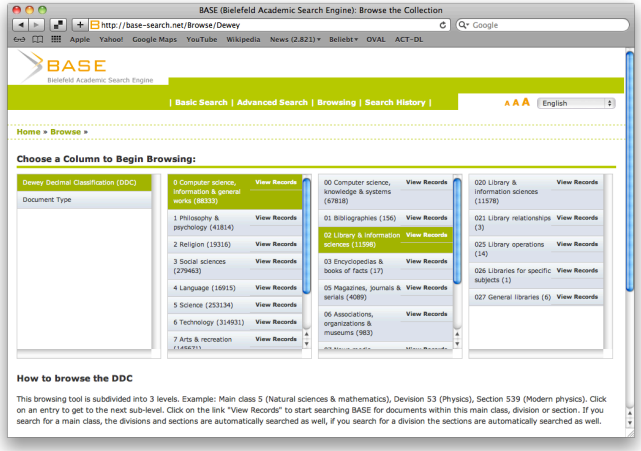
Überwachtes Lernen

Zielkategorien bekannt

Spam-Filter

DEUTSCHE NATIONAL BIBLIOTHEK

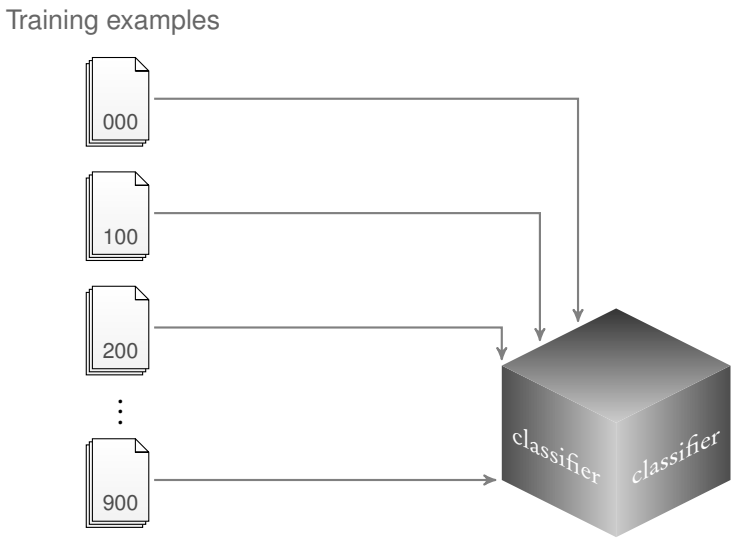
Klassifikation nach DNB-Sachgruppen



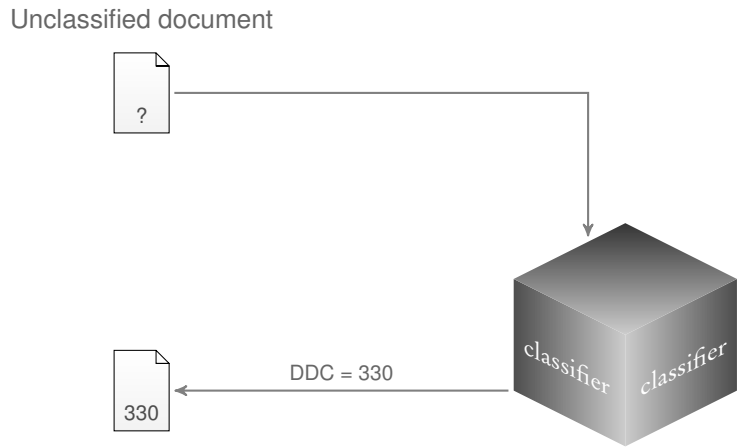
Klassifikation nach DDC

Klassifikation nach DDC

Lernphase



Applikationsphase



Task *T*: Klassifiziere beliebige Textdokumente nach der DDC

Experience *E*: DDC-klassifizierte Beispieldokumente

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

Trainingsdokumente

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

Trainingsdokumente

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

Trainingsdokumente

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

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

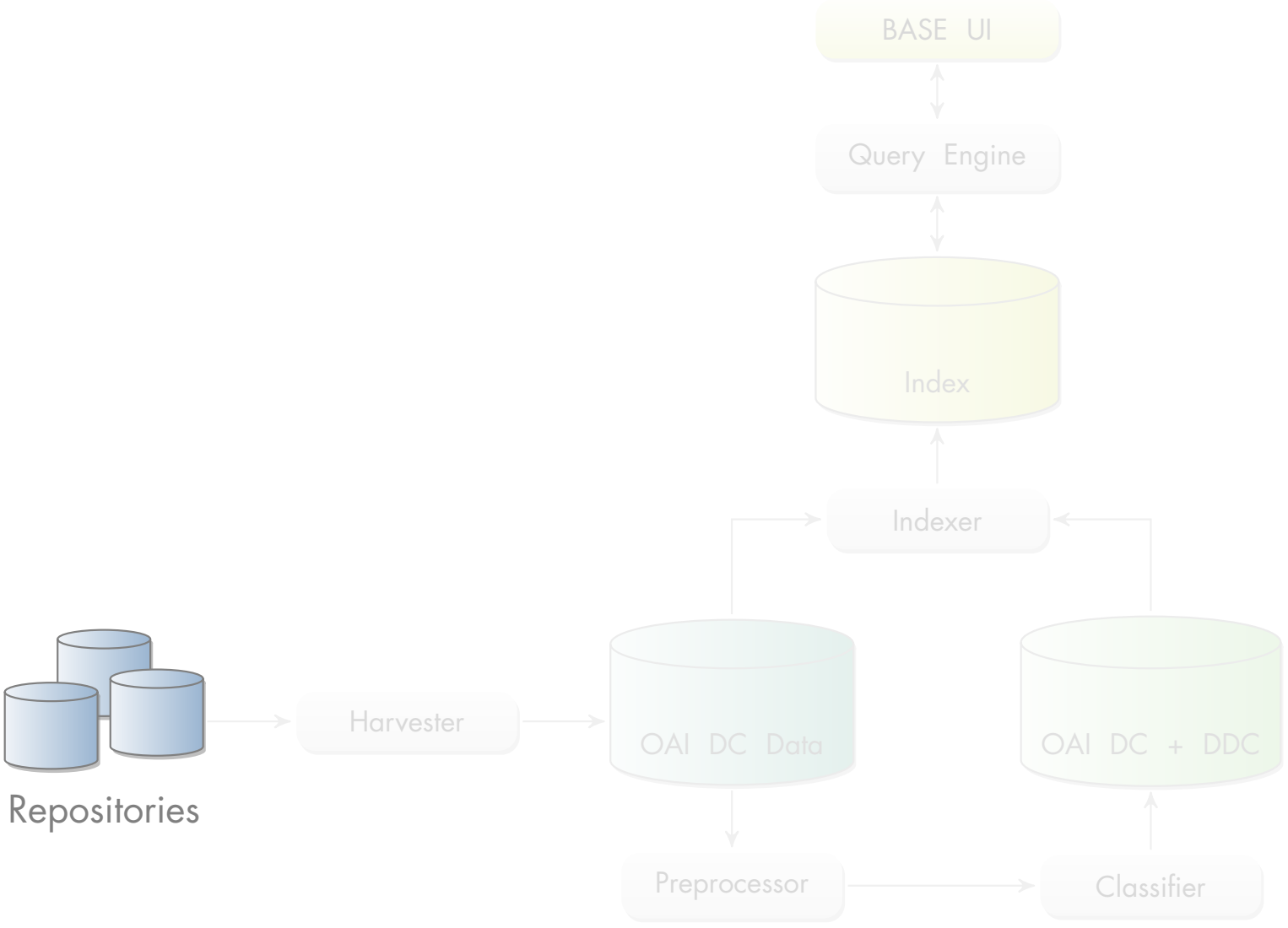
```
<record>
  ...
  <metadata>
    <oai_dc:dc xmlns:xsi="...">
      <dc:title>
        Bielefeld Academic Search Engine (BASE): an end-user oriented
        institutional repository search service
      </dc:title>
      <dc:creator>Pieper, Dirk</dc:creator>
      <dc:creator>Summann, Friedrich</dc:creator>
      <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

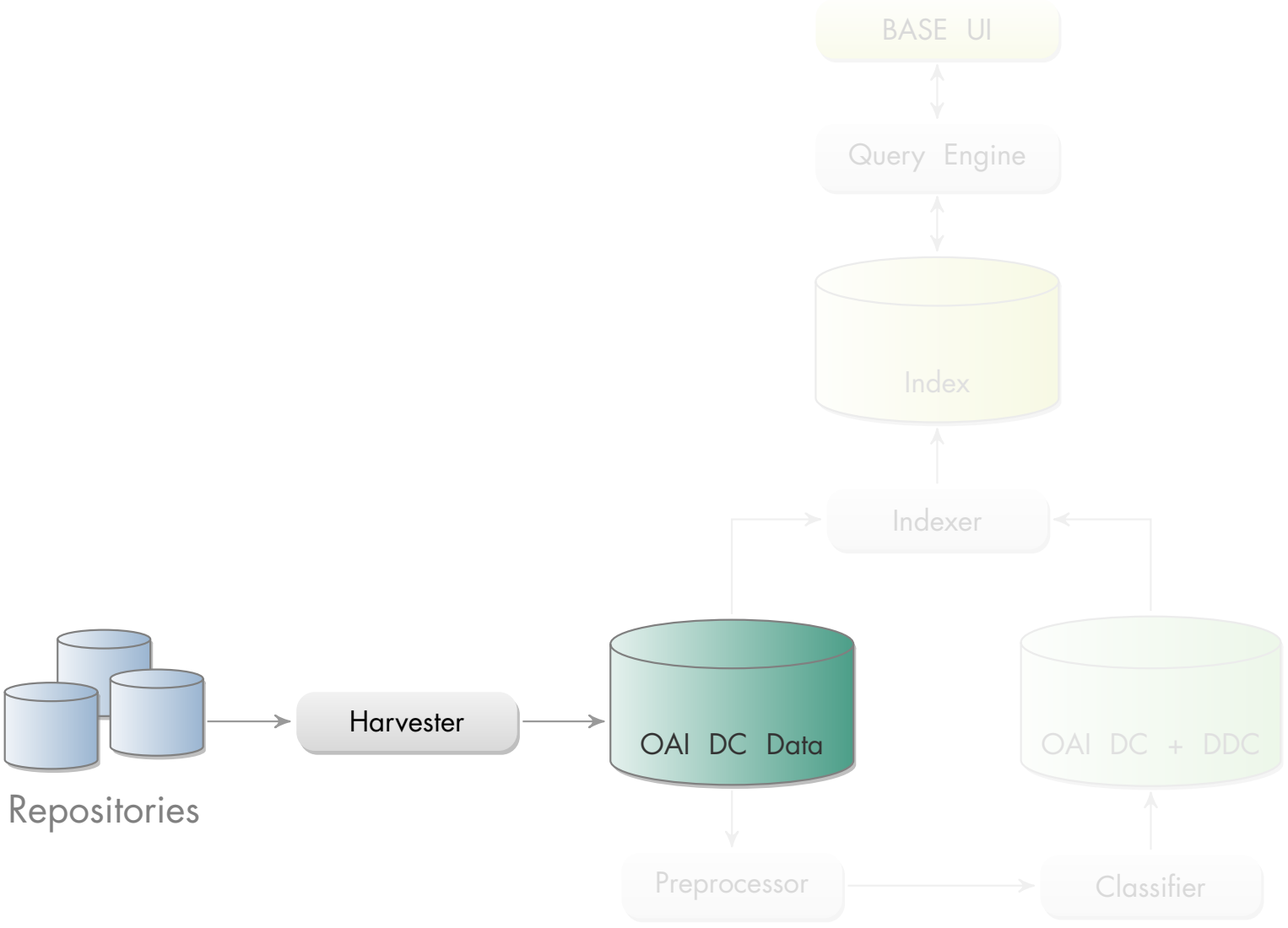
OAI-DC-Metadaten

```
<metadata>
  <oai_dc:dc xmlns:xsi="...">
    <dc:title>
      Bielefeld Academic Search Engine (BASE): an end-user orien
      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 i
      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 ove
      the functionalities of BASE and gives insight into the cha
      have to be faced when harvesting and integrating resources
      OAI servers.
    </dc:description>
```

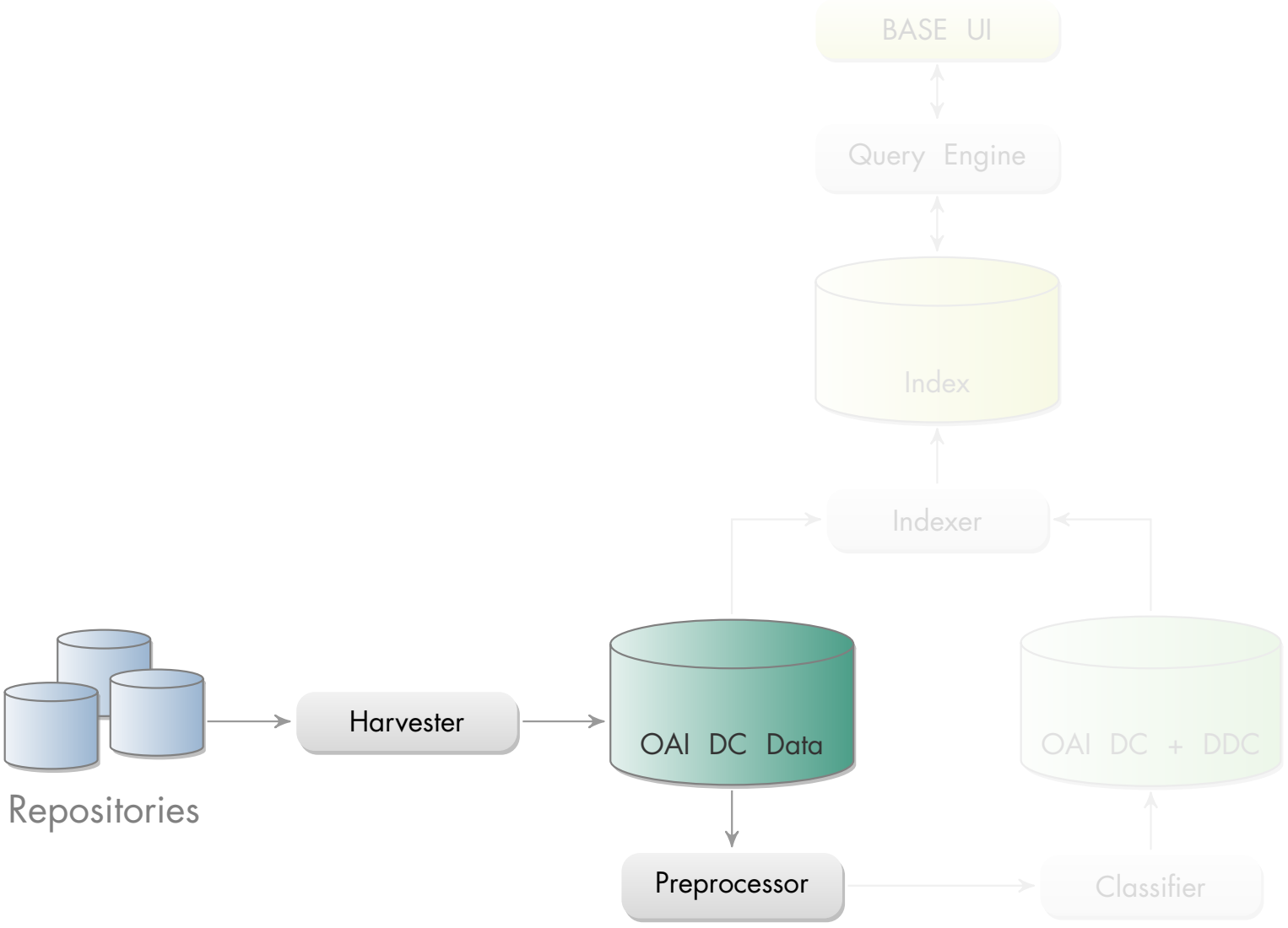
Metadatenanreicherung in BASE



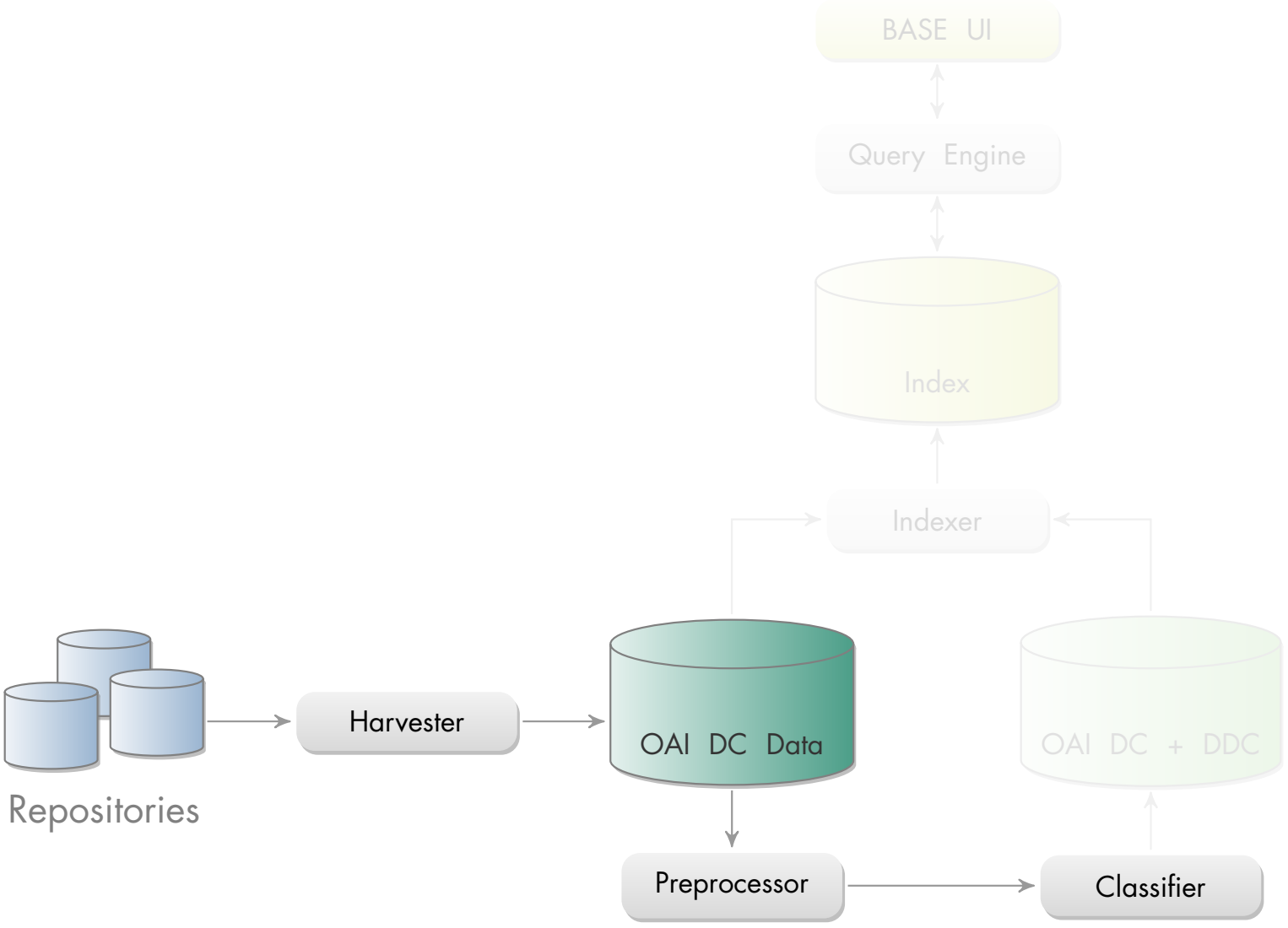
Metadatenanreicherung in BASE



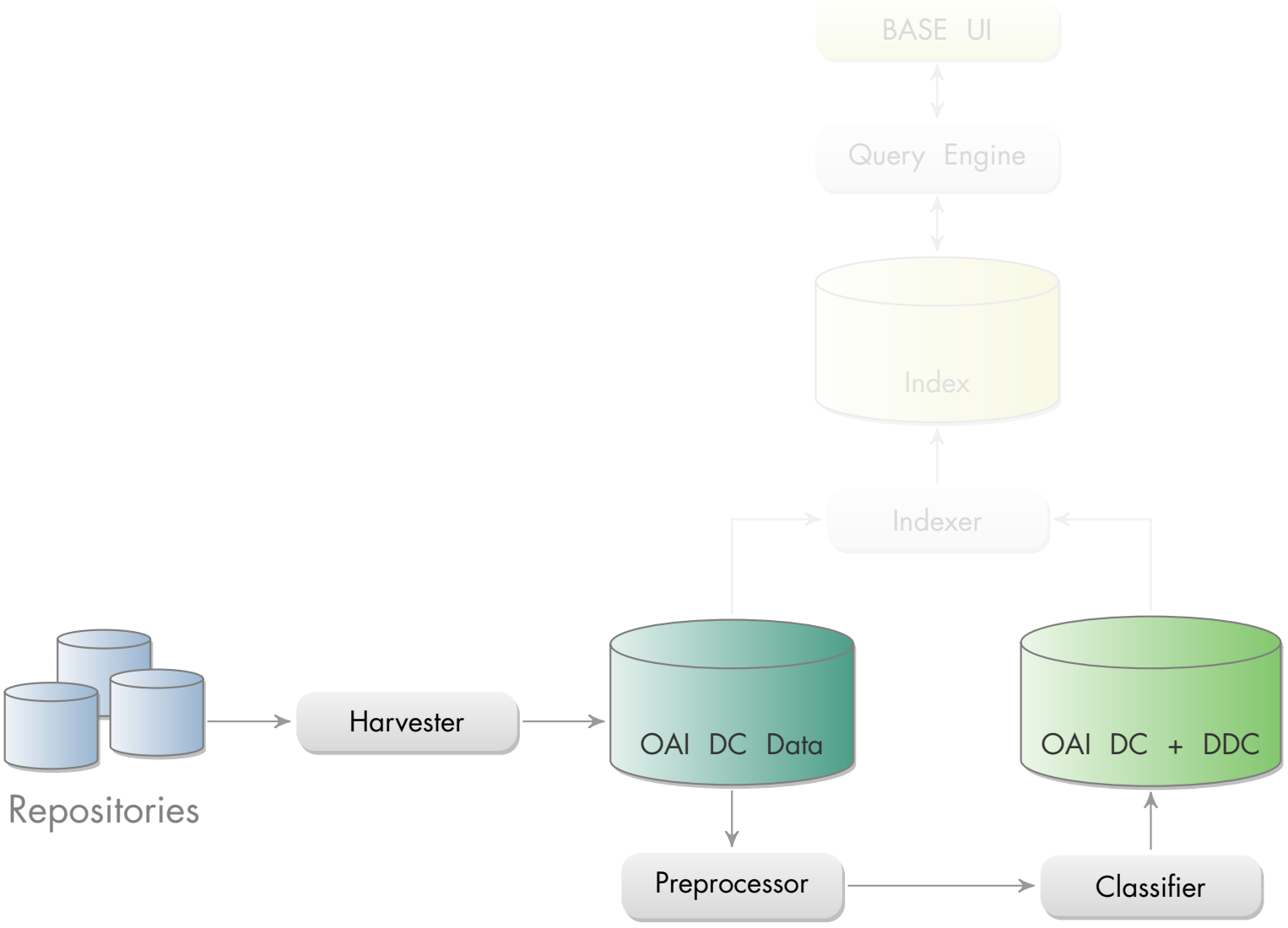
Metadatenanreicherung in BASE



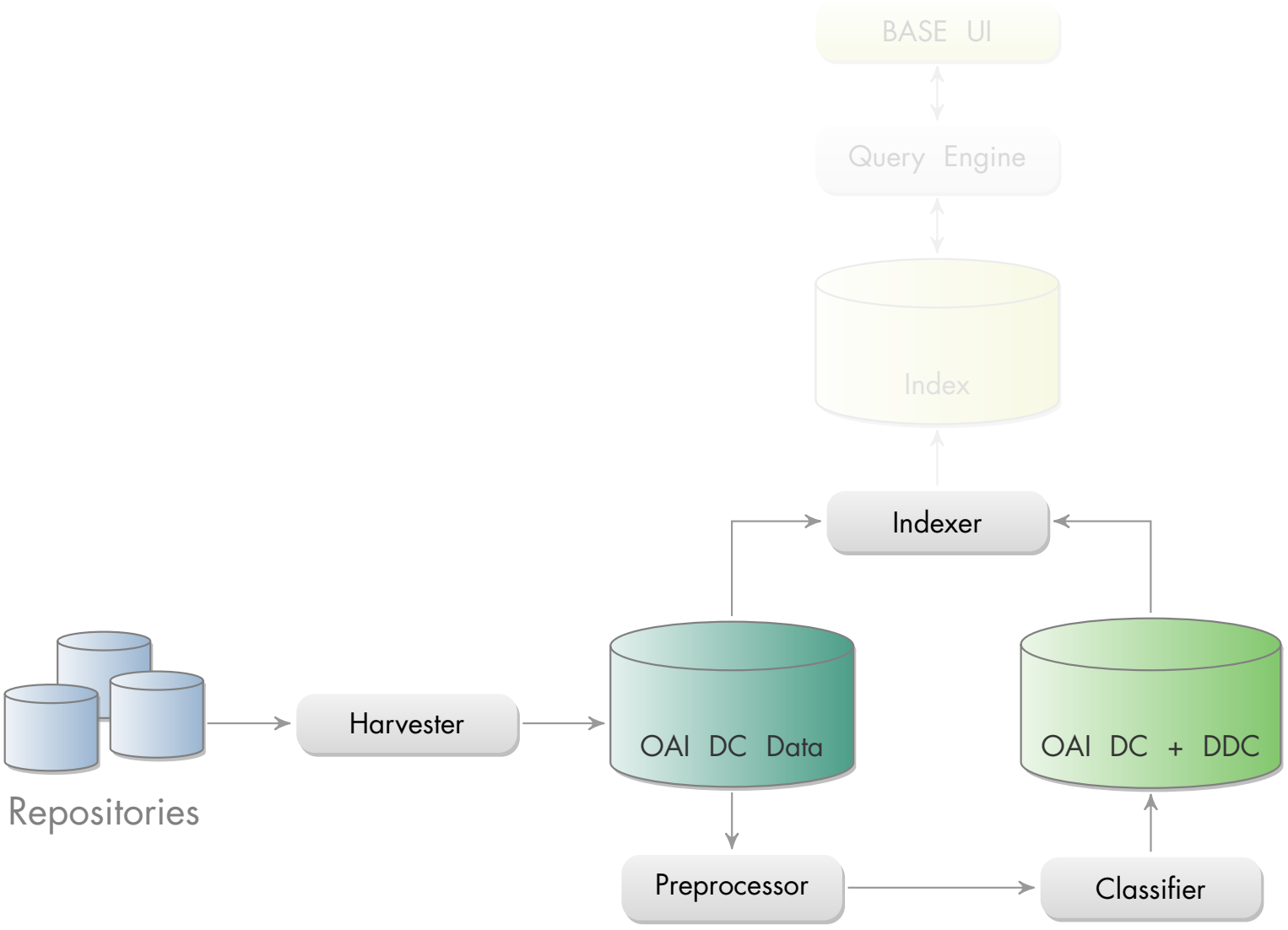
Metadatenanreicherung in BASE



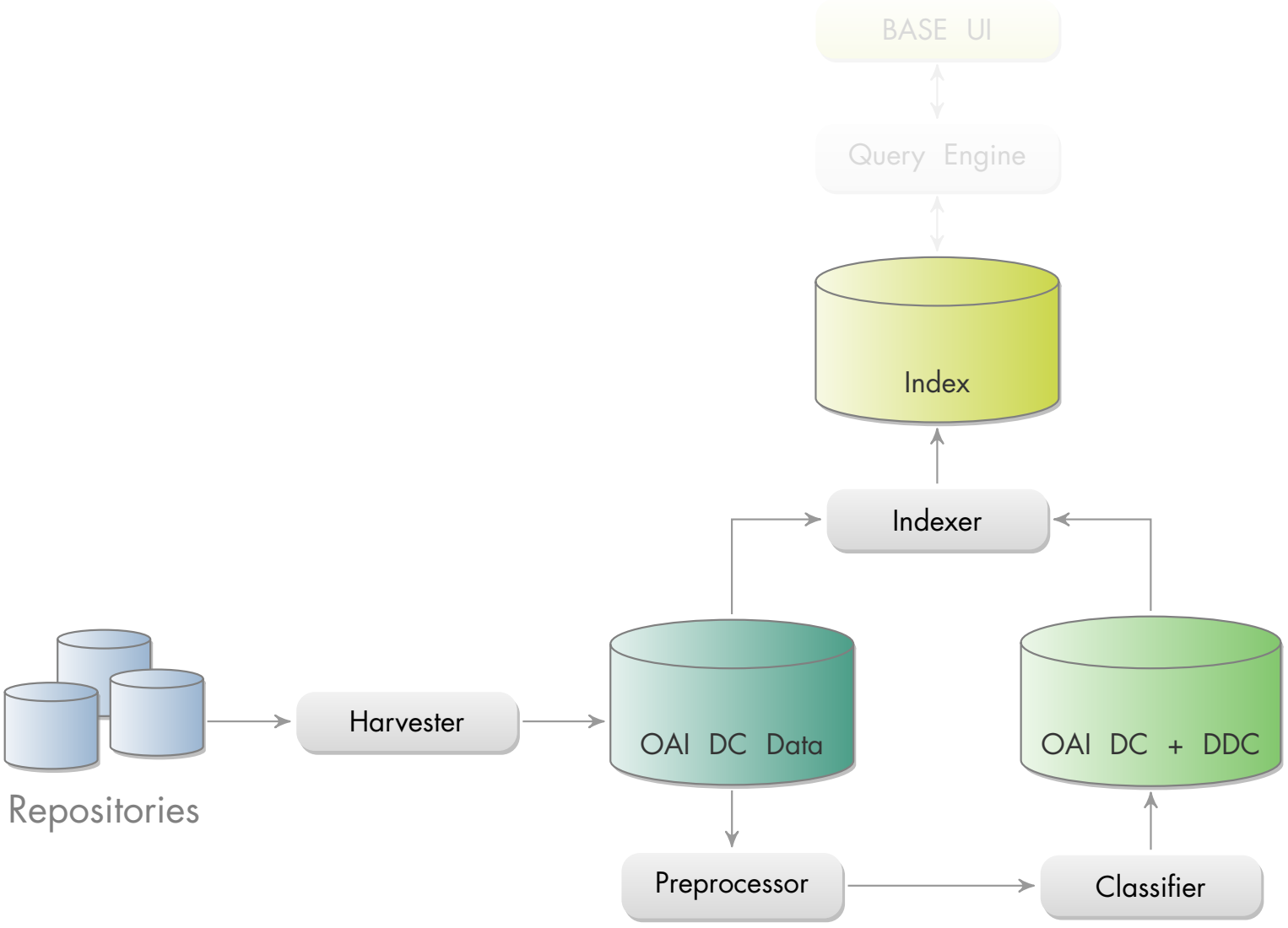
Metadatenanreicherung in BASE



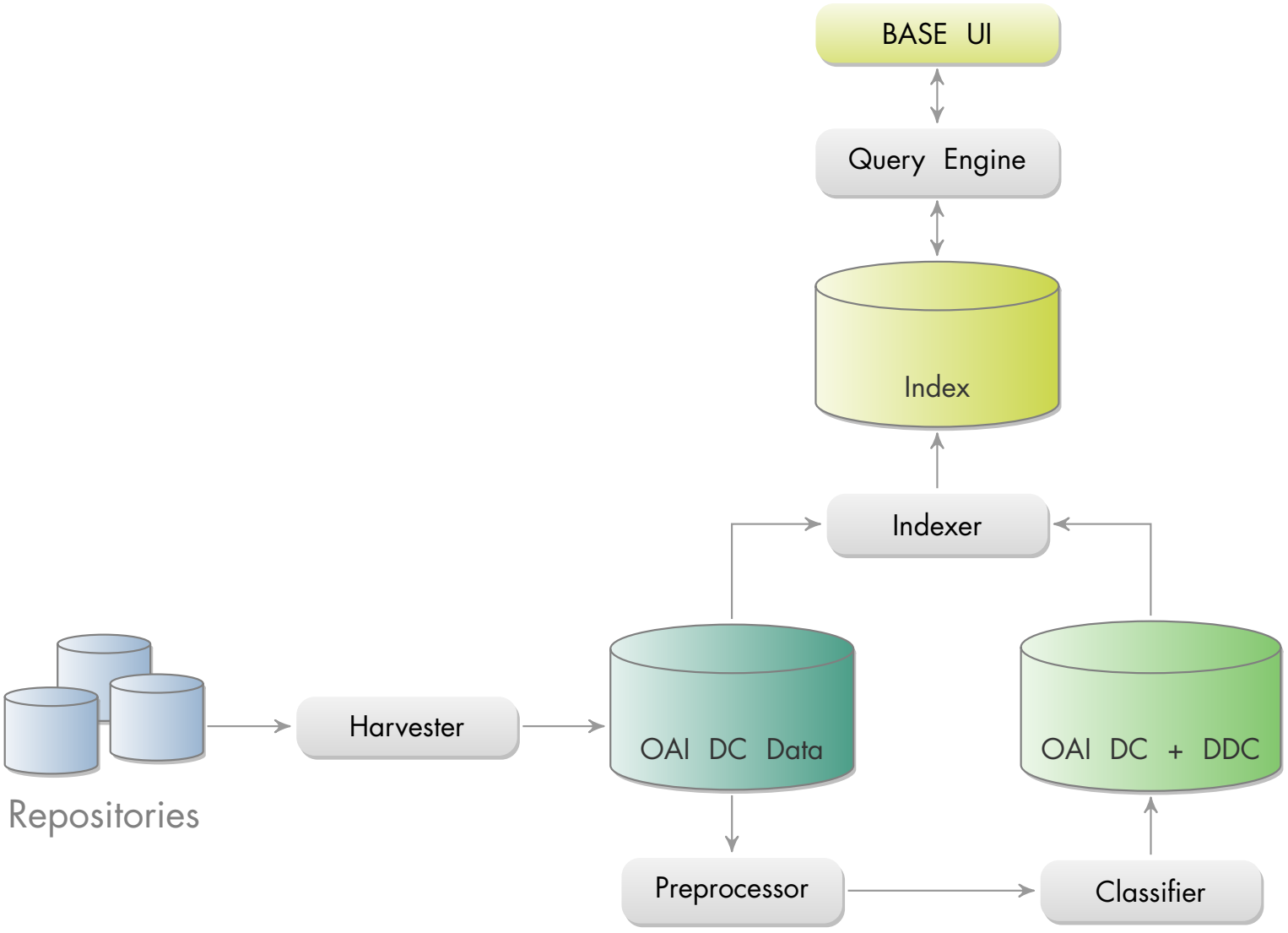
Metadatenanreicherung in BASE



Metadatenanreicherung in BASE

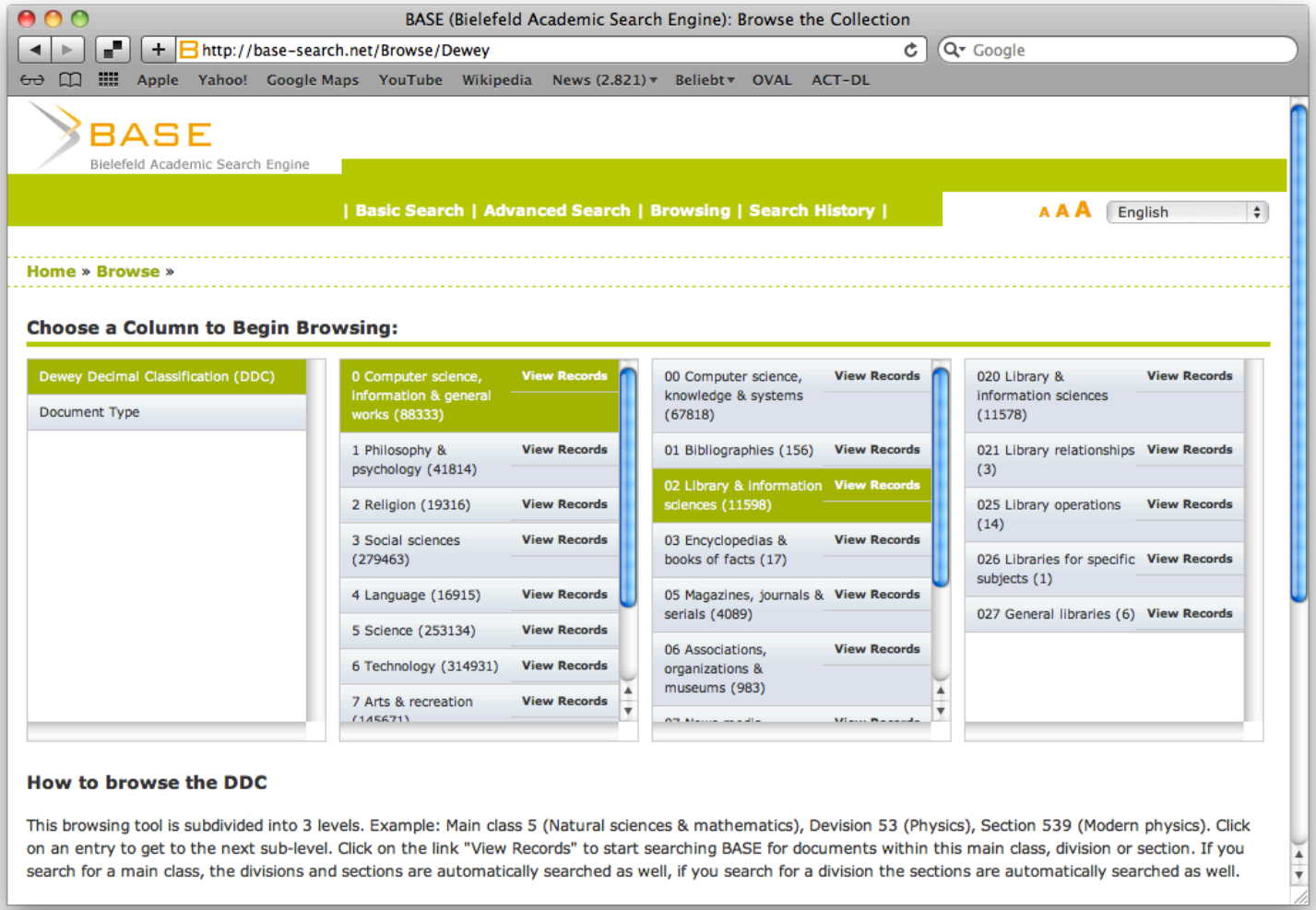


Metadatenanreicherung in BASE



BASE Browsing Interface

Entwickelt von Renata Mitrenga

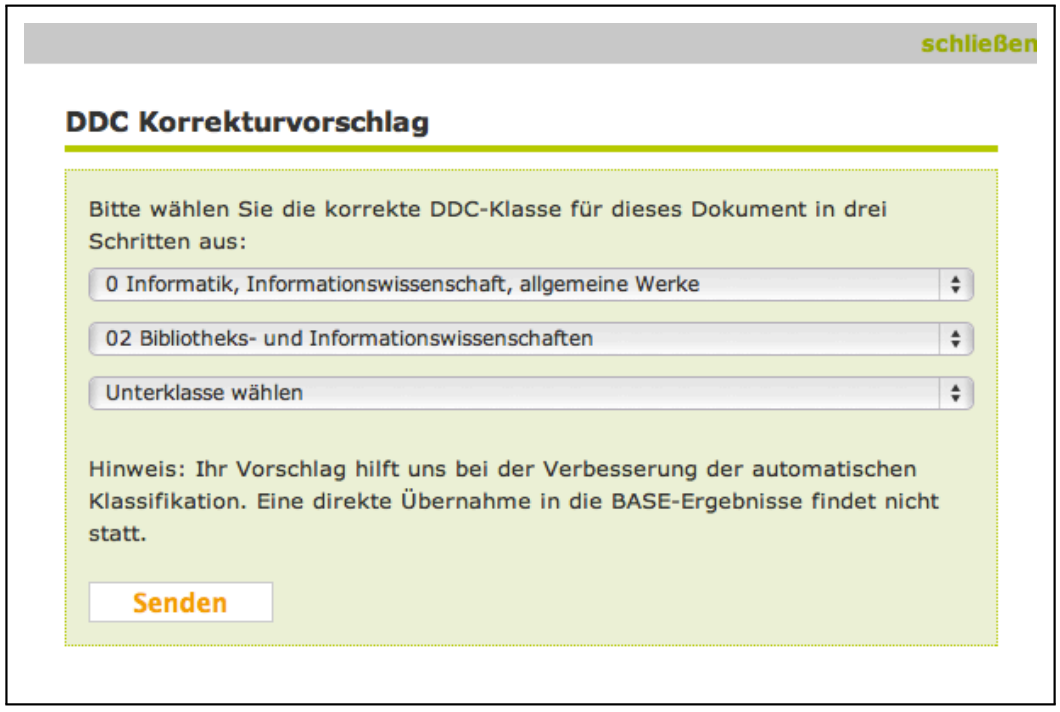


▸ Browsing aufrufen

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

Feedback-Formular

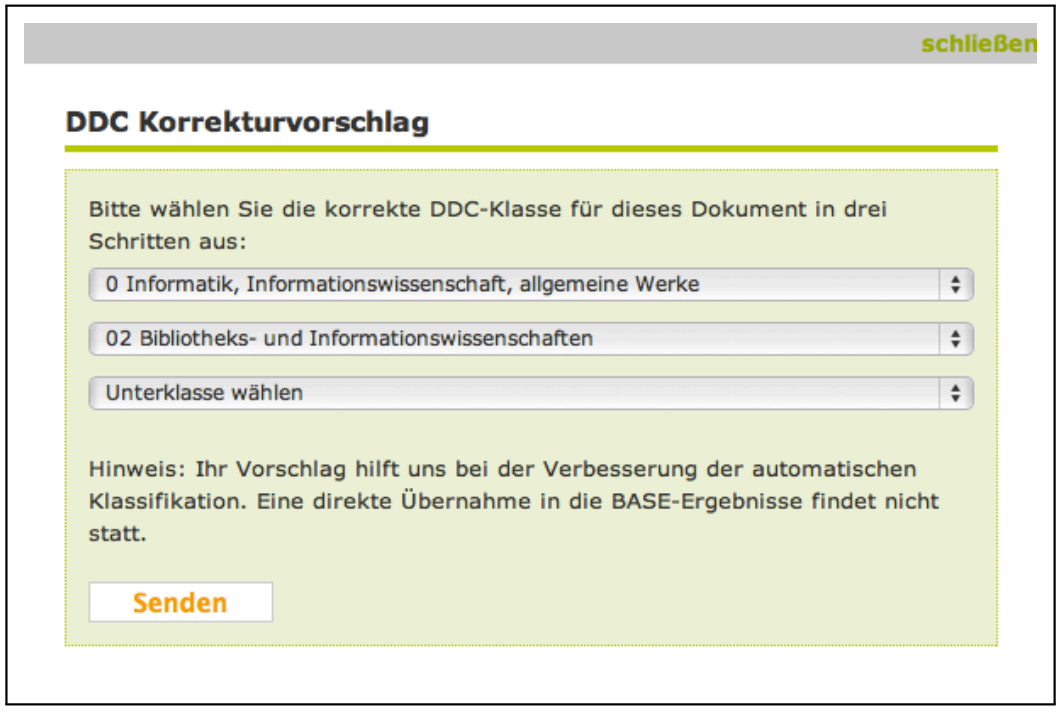
Entwickelt von Renata Mitrenga



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

Feedback-Formular

Entwickelt von Renata Mitrenga



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

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 %

Ergebnisse

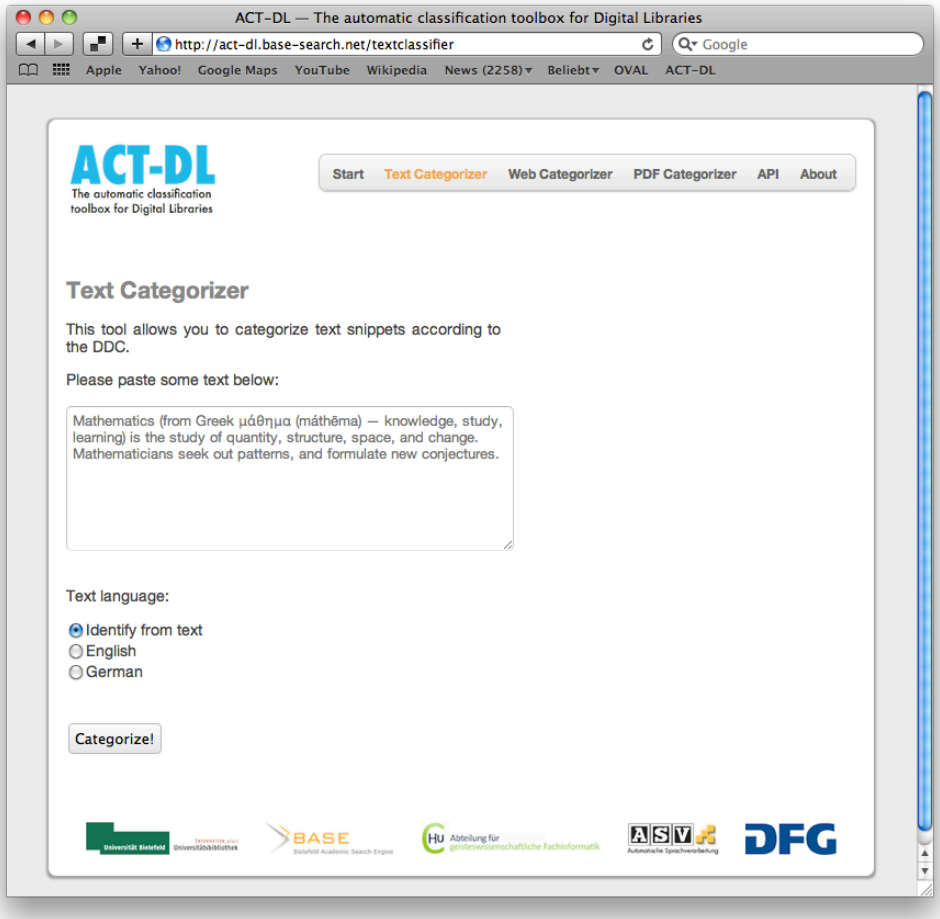
Produktive Klassifikation in BASE

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

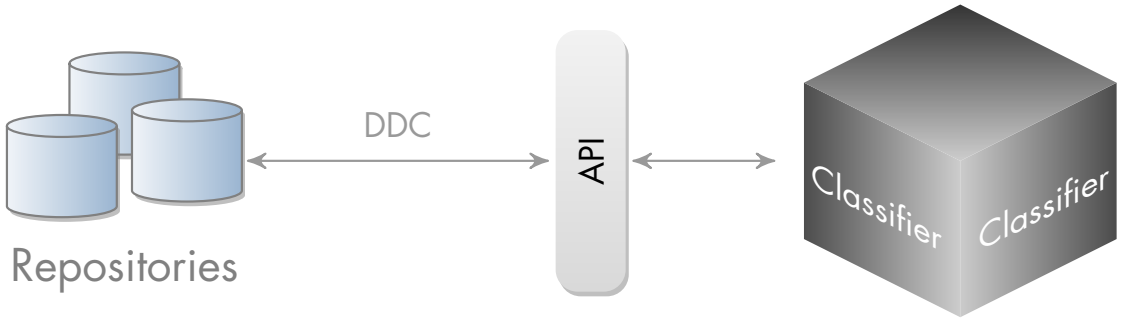


- Text Categorizer
- Web Categorizer
- PDF Categorizer

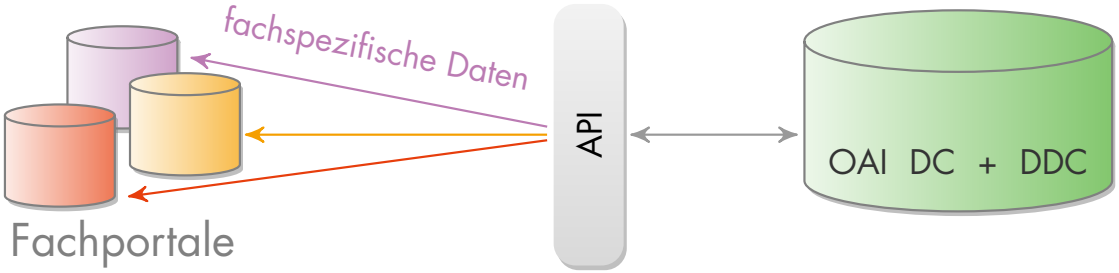
▸ Seite besuchen

Universität Bielefeld
Information.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

Nachnutzung des Klassifikators



Nachnutzung klassifizierter Metadaten



Nachnutzung: Klassifikator

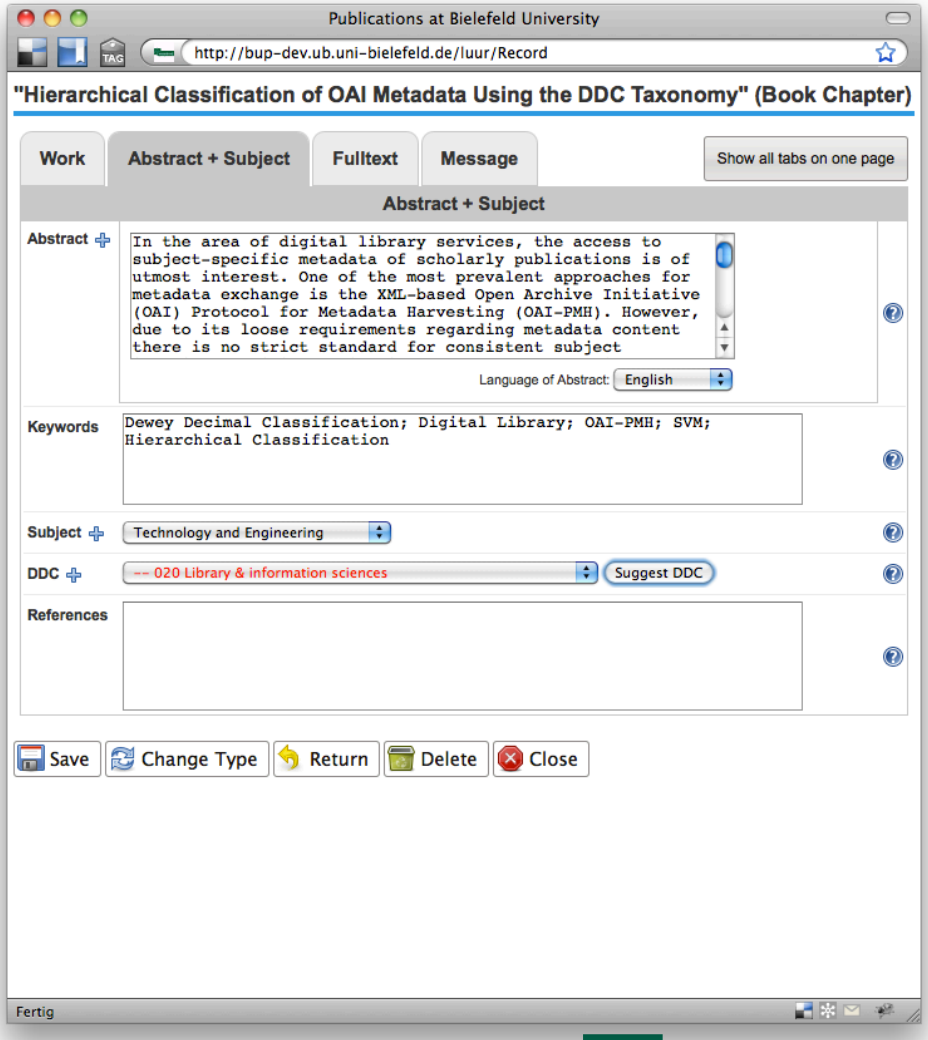
API

```
<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 &amp;amp; recreation" confidence="0.0509034274529"/>
    <DDC number="0" heading="Computer science ..." confidence="..."/>
    <DDC number="9" heading="History &amp;amp; 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 &amp;amp; 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 &amp;amp; geology" confidence="0.00250678827777"/>
    <DDC number="52" heading="Astronomy" confidence="0.00108935568839"/>
    <DDC number="56" heading="Fossils &amp;amp; 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 &amp;amp; applied mathematics" confidence="..."/>
  </result>
</results>
```

▸ API-Dokumentation

Nachnutzung des Klassifikators in PUB

Vorschlagsystem für die Metadatenerfassung



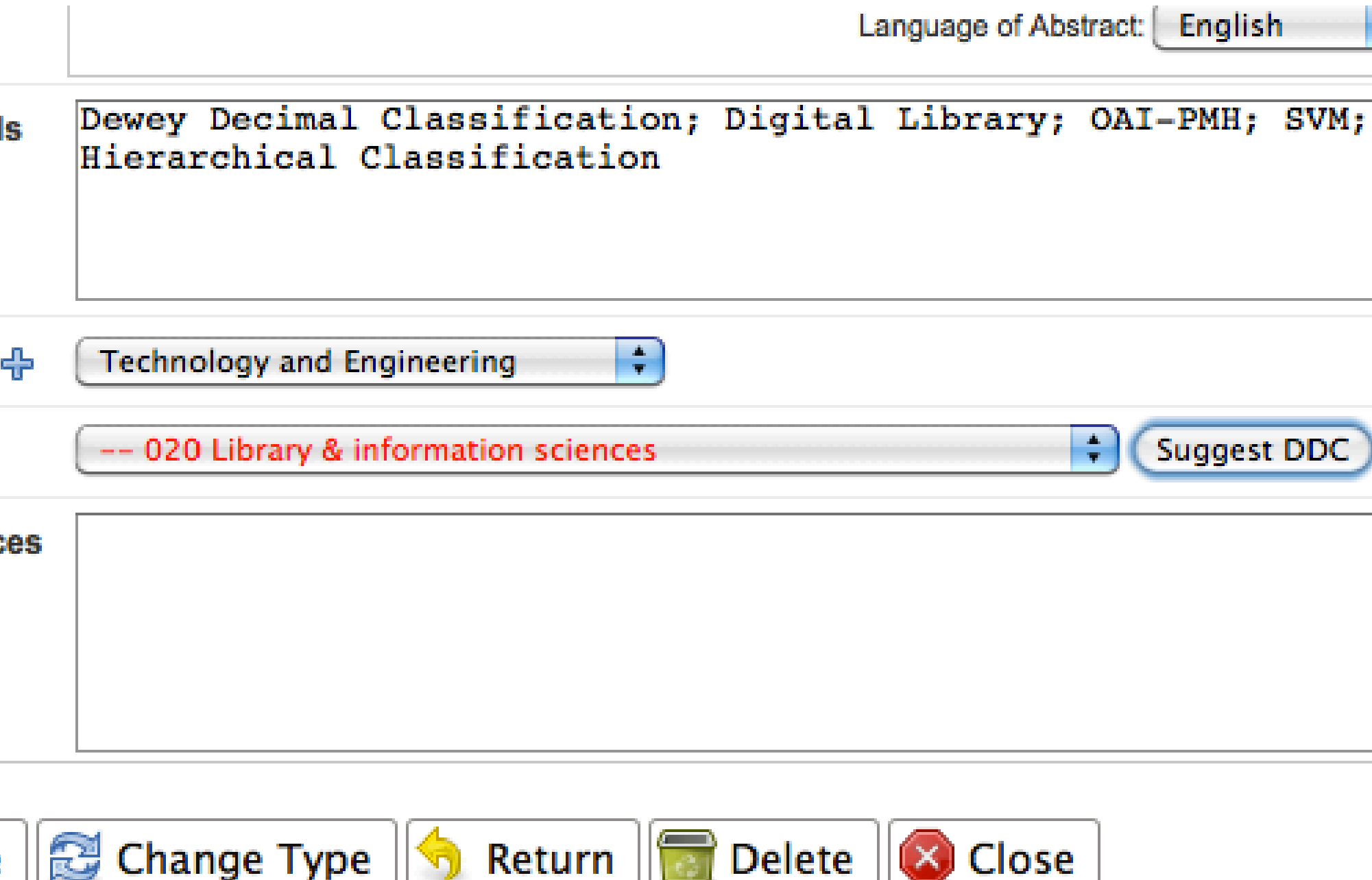
Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

BASE Bielefeld Academic Search Engine

Nachnutzung des Klassifikators in PUB

Vorschlagsystem für die Metadatenerfassung



Nachnutzung: Klassifizierte Metadaten

- 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**

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

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

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

Vielen Dank für die Aufmerksamkeit!

Universität Bielefeld | Universitätsbibliothek
Universitätsstr. 25
D-33615 Bielefeld
☎ +49 521 106-2546

✉ Mathias.Loesch@uni-bielefeld.de

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
BASE Bielefeld Academic Search Engine

Literatur

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.