Automatische Sacherschließung elektronischer Dokumente nach DDC

Mathias Lösch

Universitätsbibliothek Bielefeld

Mathias.Loesch@uni-bielefeld.de

Gegenwart und Zukunft der Sacherschließung
7. Oktober 2011

Hintergrund

- DFG-Projekt »Automatische Anreicherung von OAI-Metadaten«
- Förderung Oktober 2009–September 2011
- Partner:

Universitätsbibliothek Bielefeld/BASE

Text Technology Lab
Unversität Frankfurt/Main

Institut für automatische Sprachverarbeitung
Universität Leipzig

Agenda

1. Motivation
2. Automatische Klassifikation
3. Ergebnisse
4. Zusammenfassung

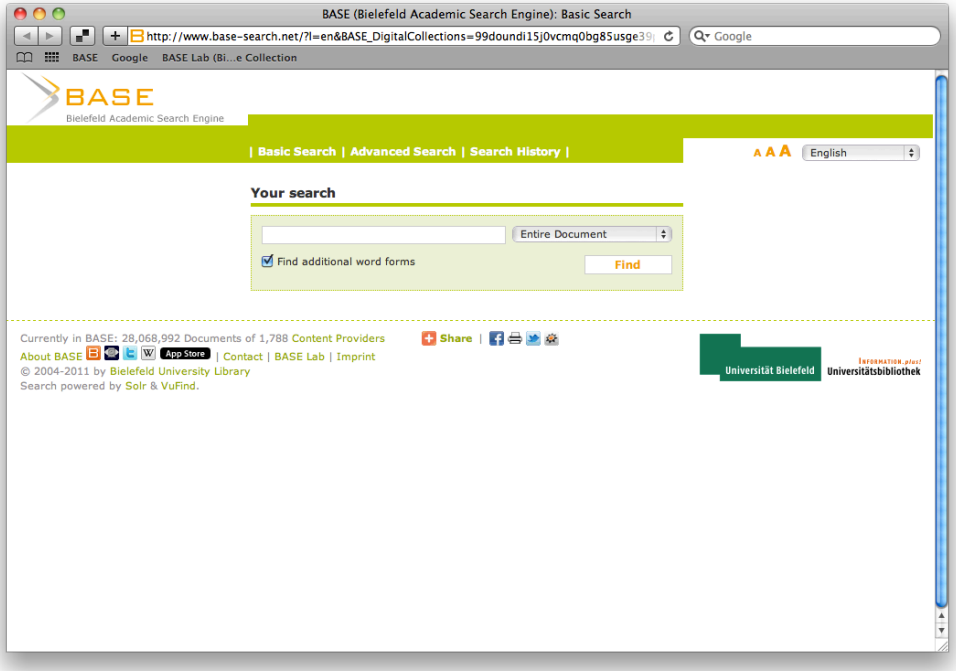
1 **Motivation**

2 Automatische Klassifikation

3 Ergebnisse

4 Zusammenfassung

Motivation



BASE
Bielefeld Academic Search Engine

- Wissenschaftliche Suchmaschine
- Nachweis von > 31 Mio Dokumenten
- Aggregation der Inhalte von > 2.000 Dokumentenservern

Motivation

BASE-Datenquellen

```
<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>
```

- OAI-PMH-Protokoll
- OAI-DC-Metadaten

Motivation

Eigenschaften von OAI-DC-Daten:

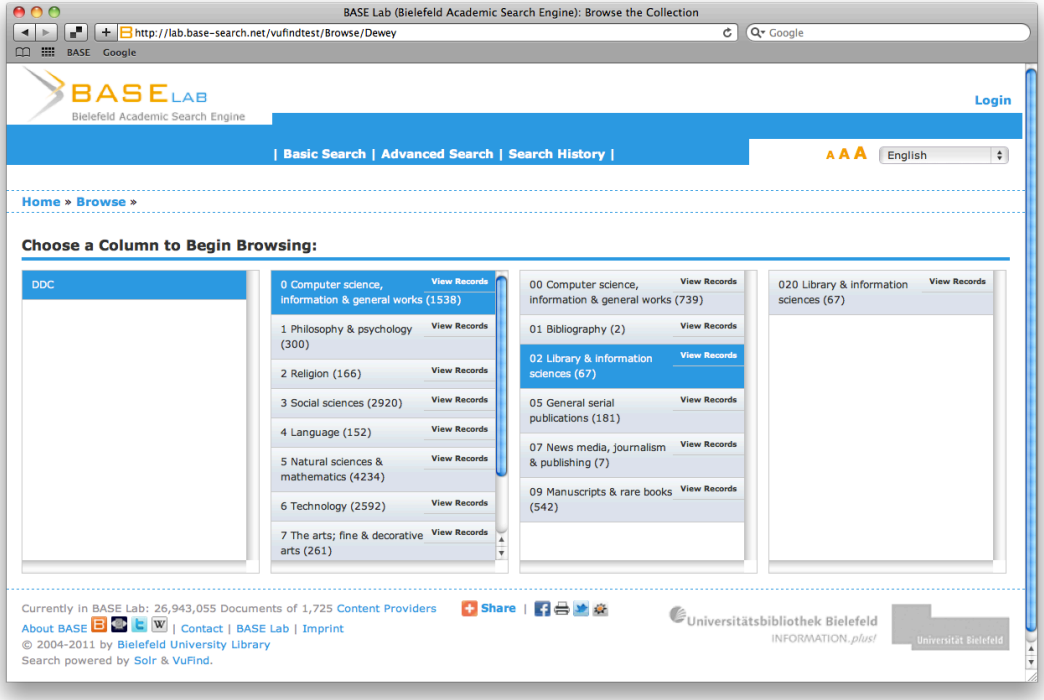
- enthalten selten Sacherschließungsinformationen
- sehr heterogene Erschließungsformen:
  - Fach- und Universalklassifikationen
  - Kontrollierte Vokabulare
  - Freie Schlagwörter
  - …

Motivation

BASE DDC-Browsing

Intellektuelle Erschließung

- Verbreitetste Klassifikation: DDC
- 443.249 Dokumente
- ≈ 1,41% der Datenbasis



Motivation
Fachportale

- Trend zu disziplinspezifischen Portallösungen
  - PubMed (Life sciences)
  - arXiv.org (Physik, Informatik, …)
  - EconBiz (Wirtschaft)
  - ELIS (Bibliotheks-/Informationswissenschaft
  - SSOAR (Sozialwissenschaft)
  - …
- Interesse an automatischer Extrahierung fachlicher Subsets aus der BASE-Datenbasis

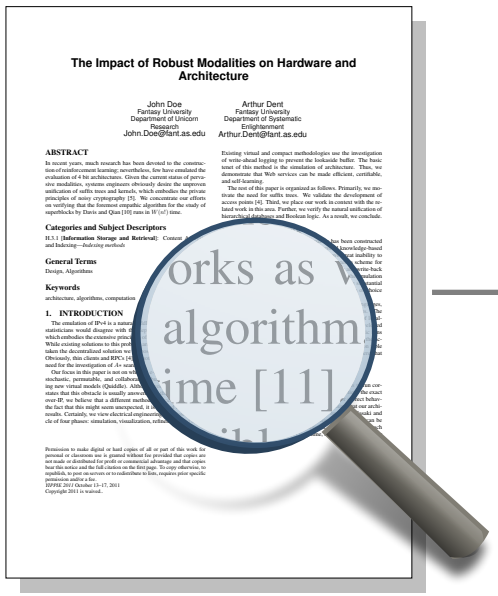
1 Motivation

2 **Automatische Klassifikation**

3 Ergebnisse

4 Zusammenfassung

Automatische Klassifikation



MEDIZIN

BIOLOGIE

LITERATURWISSENSCHAFT

MATHEMATIK

INFORMATIK

PHYSIK

GESCHICHTE

POLITIKWISSENSCHAFT

SOZIOLOGIE

Universität Bielefeld

Information.plus! Universitätsbibliothek

BASE Bielefeld Academic Search Engine

DFG

Referenzontologie: Dewey Dezimalklassifikation

Vorteile der DDC

- Universalklassifikation
- international starke Verbreitung (~200.000 Bibliotheken weltweit)
- Numerische Notation/Dezimalstruktur: Sprachunabhängige Kodierung der Klassen, auf-/absteigende Traversierung durch Trunkierung/Expansion der Nummern einfach möglich
- Durch Empfehlung von DINI in der deutschen Repository-Landschaft meist-verwendete Klassifikation

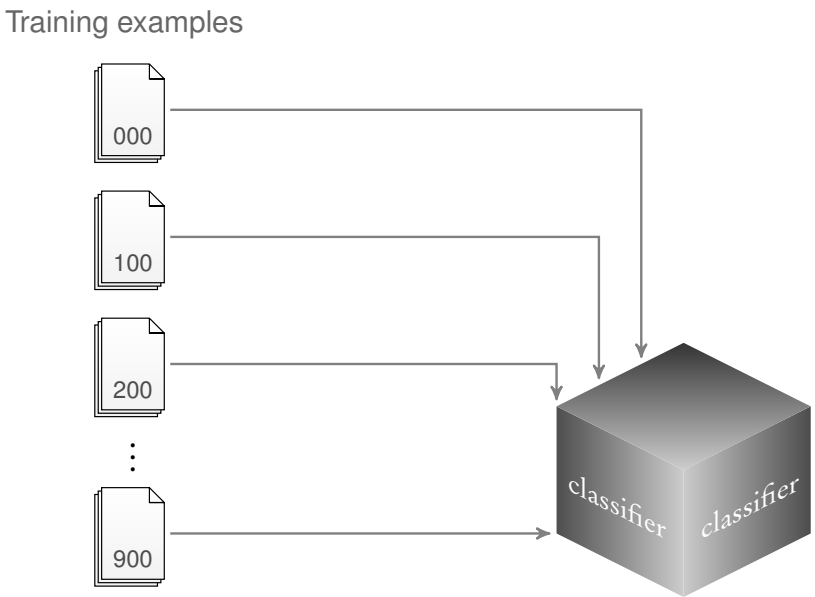
Automatischer Klassifikator

Maschinelles Lernen (Sebastiani, 2002)

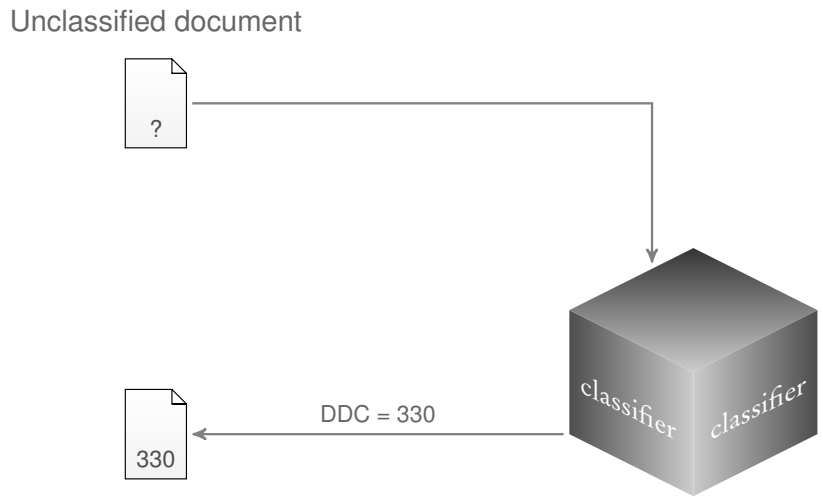
- Automatische Generierung eines Klassifikators aus Beispieldokumenten
- Lernen von Klassen durch extensionale Beschreibung (= Aufzählung von Beispielen)

Automatischer Klassifikator

Lernphase



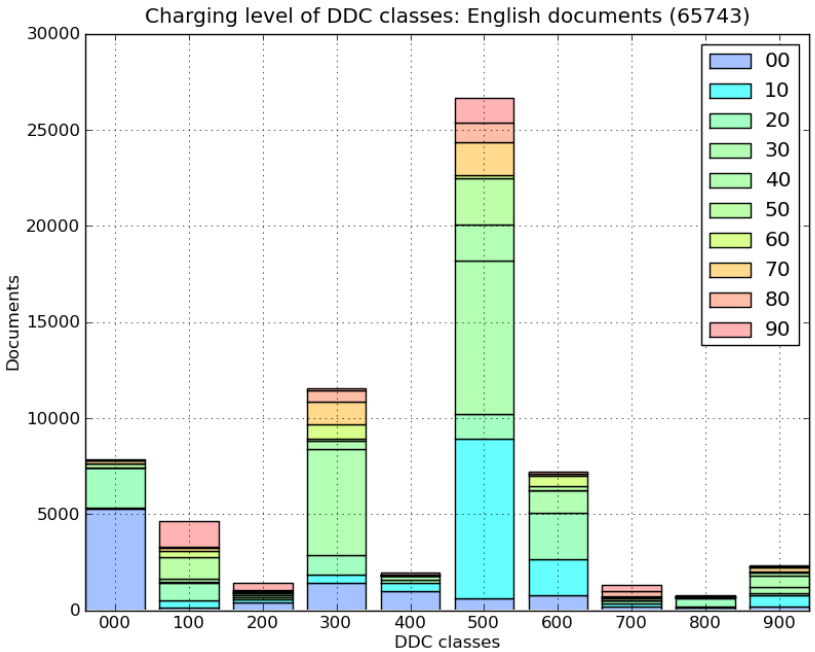
Applikationsphase

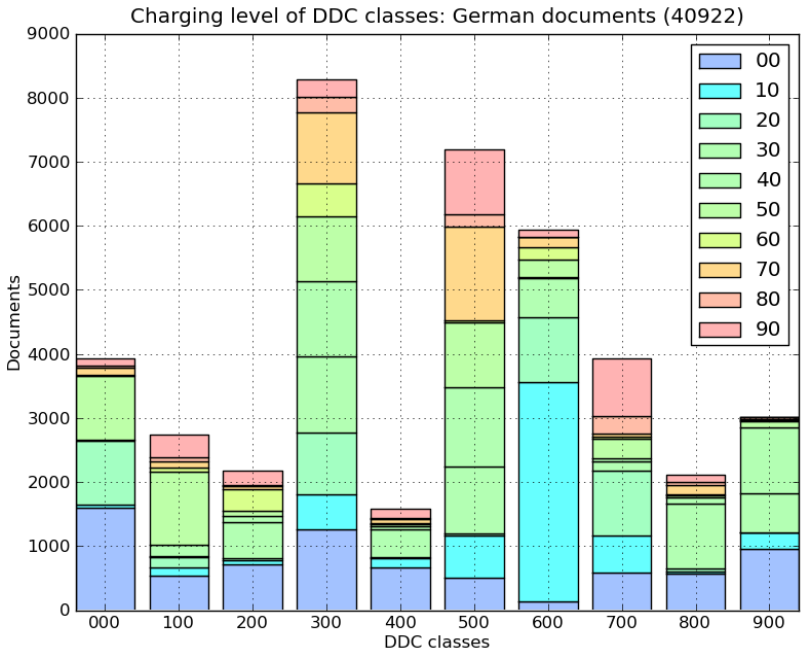


Trainingsdokumente (Lösch et al., 2011)

- Konstruktion eines DDC-kategorisierten Textkorpus aus der BASE-Datenbasis
- OAI DC Metadaten (+ Volltexte)
- Deutsch und Englisch
- semi-automatische Vergabe von DDC-Nummern über Konkordanzen
- ~100.000 Dokumente

Korpuserstellung





- Schiefe Verteilung der Dokumente über die DDC-Klassen
- Wenig Beispieldokumente in den Geisteswissenschaften
- Dokumentakquise ab der dritten DDC-Ebene (1.000 Klassen) extrem aufwändig mangels guter Sacherschließungsinformationen in den Trainingsdaten

1 Motivation

2 Automatische Klassifikation

3 **Ergebnisse**

4 Zusammenfassung

Ergebnisse: Klassifikationsgenauigkeit

Evaluationsmethode

- Cross-Validation
- Leave-One-Out (LOO)
- Recall, Precision und F1-Wert

Ergebnisse: Klassifikationsgenauigkeit

DDC	Recall	Precision	F-Score
0	0.854	0.937	0.894
1	0.876	0.912	0.893
2	0.865	0.904	0.884
3	0.773	0.856	0.813
4	0.867	0.934	0.899
5	0.870	0.907	0.888
6	0.795	0.884	0.837
7	0.722	0.775	0.748
8	0.717	0.809	0.760
9	0.629	0.793	0.702
Overall	0.797	0.871	0.832

Tabelle: Ergebnisse auf der Ebene der Hauptklassen (Sprache: Englisch; Korpusgröße ~15.000 Dokumente)

Ergebnisse: Klassifikationsgenauigkeit

DDC	Recall	Precision	F-Score
30	0.636	0.901	0.746
31	0.894	0.905	0.900
32	0.764	0.884	0.820
33	0.882	0.955	0.917
34	0.772	0.855	0.811
36	0.816	0.876	0.844
37	0.920	0.883	0.901
38	0.825	0.924	0.872
39	0.859	0.971	0.912
Overall	0.819	0.906	0.858

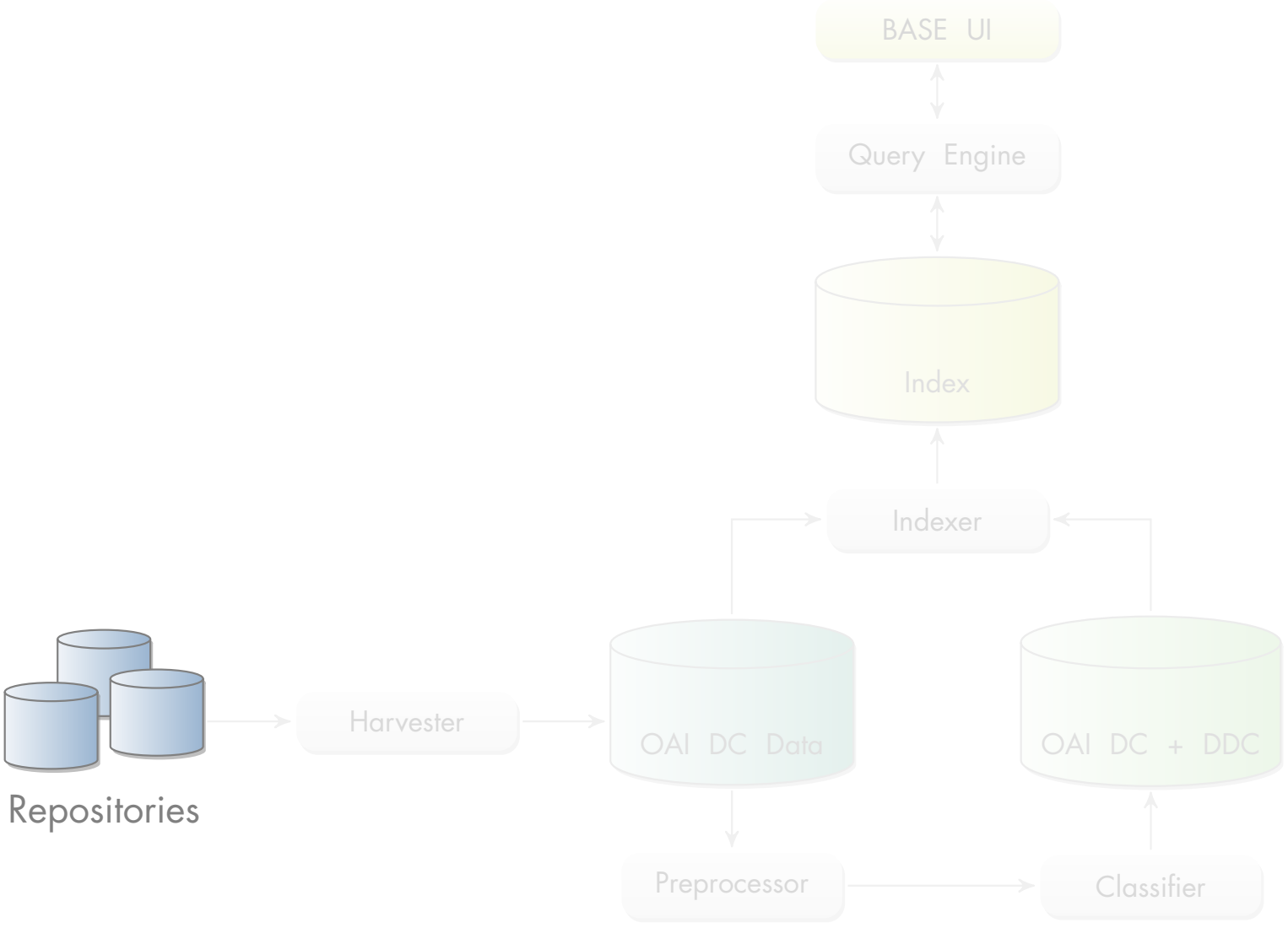
Tabelle: Ergebnisse auf der zweiten Ebene in der Division 300 Sozialwissenschaften (Sprache: Englisch; Korpusgröße ~5.000 Dokumente)

Ergebnisse: Klassifikationsgenauigkeit

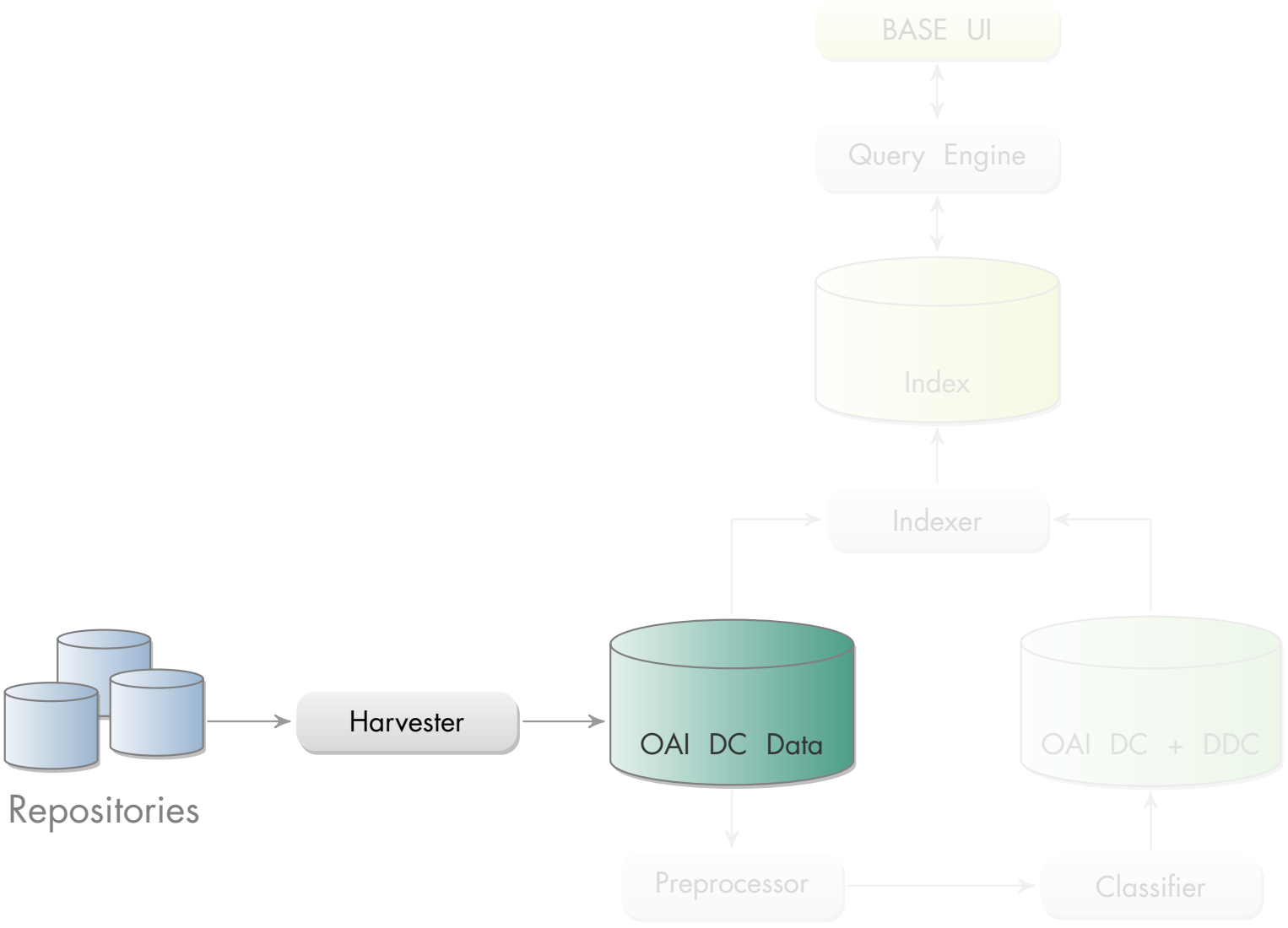
DDC	Recall	Precision	F-Score
510	0.544	0.889	0.675
511	0.352	0.694	0.467
512	0.152	0.555	0.239
514	0.247	0.667	0.361
515	0.159	0.617	0.252
516	0.186	0.567	0.281
518	0.350	0.659	0.457
519	0.382	0.672	0.487
Overall	0.297	0.665	0.402

Tabelle: Ergebnisse auf der dritten Ebene in der Sektion 510 Mathematik (Sprache: Englisch; Korpusgröße ~500 Dokumente)

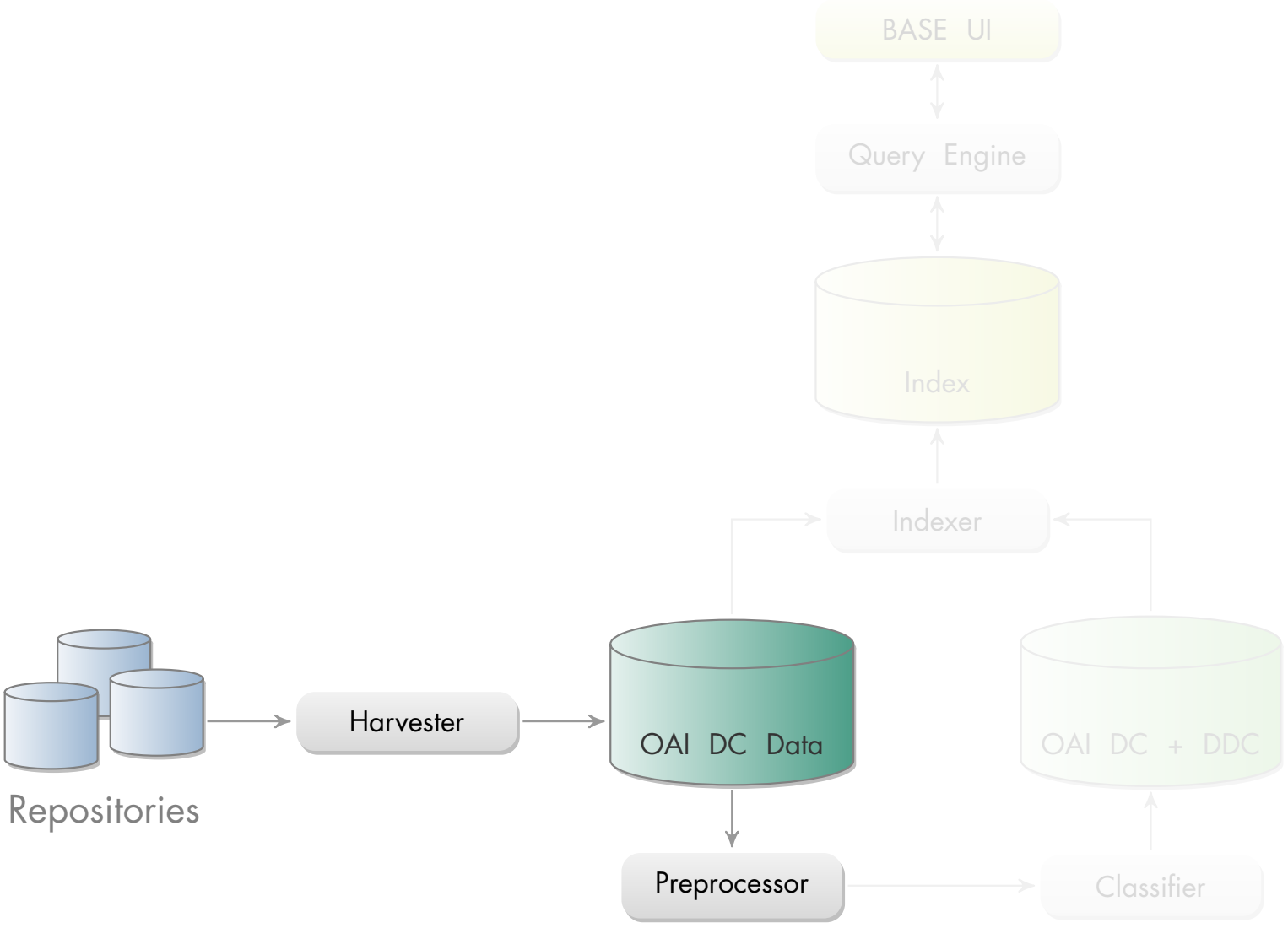
Metadatenanreicherung in BASE



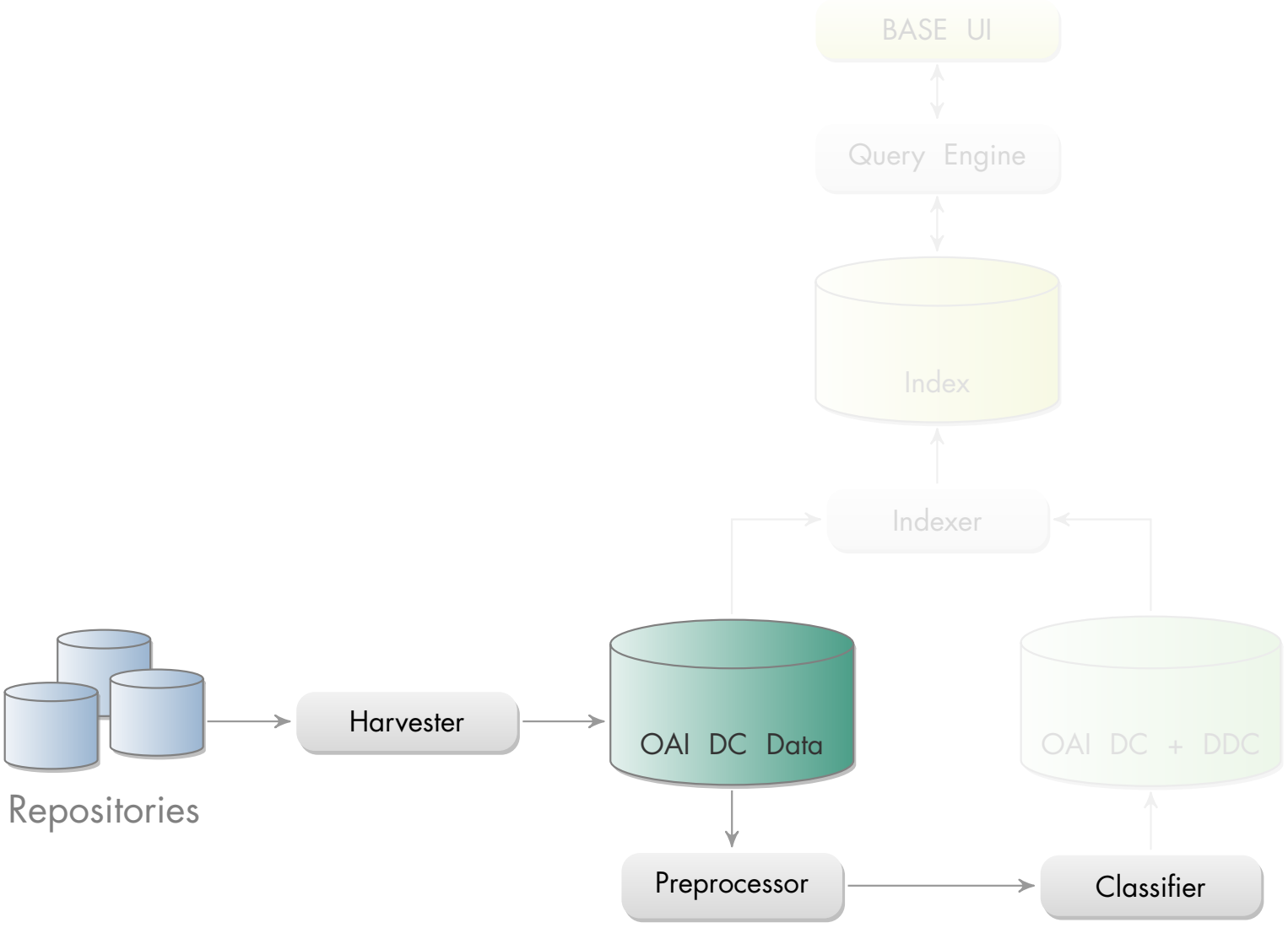
Metadatenanreicherung in BASE



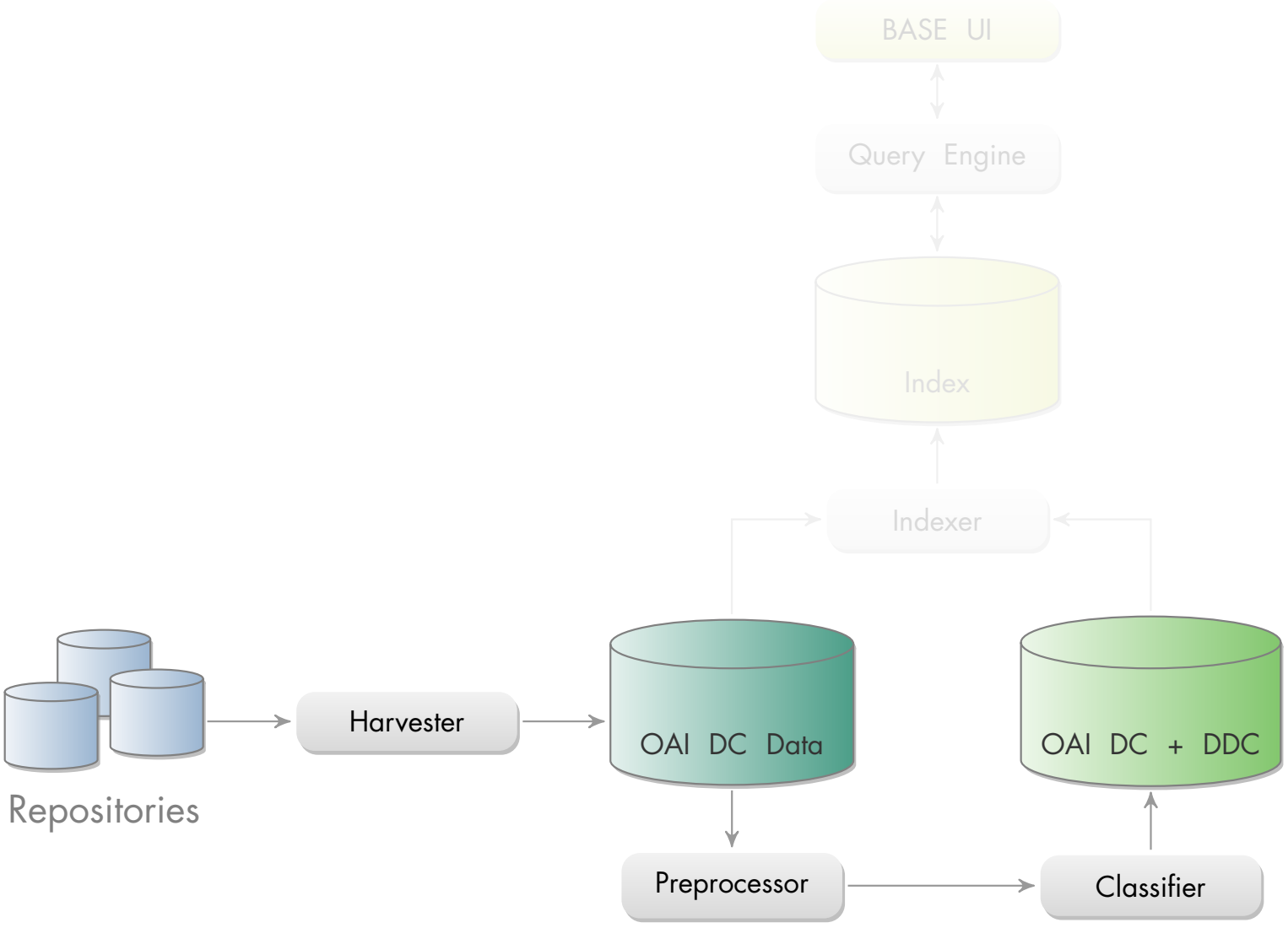
Metadatenanreicherung in BASE



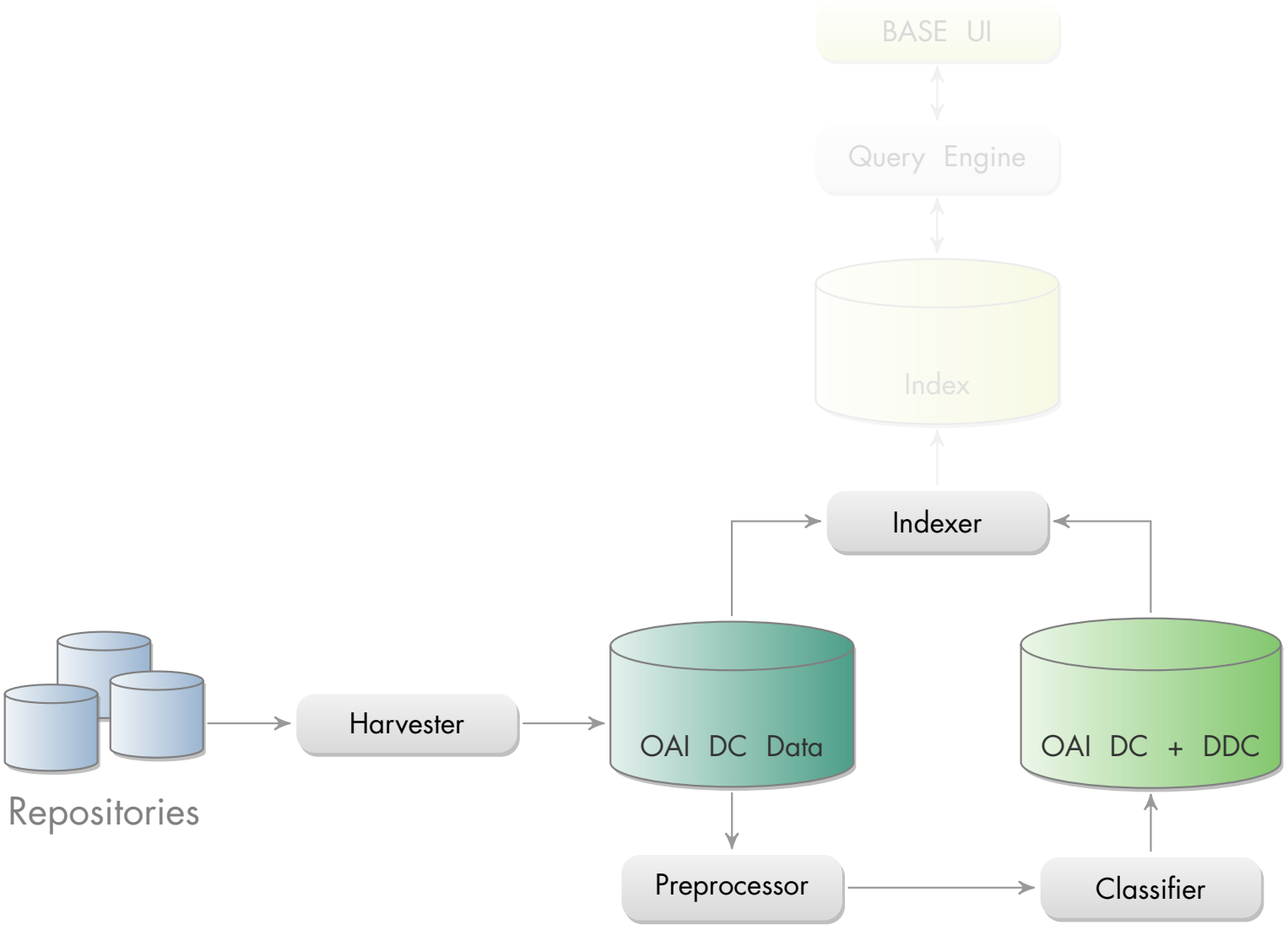
Metadatenanreicherung in BASE



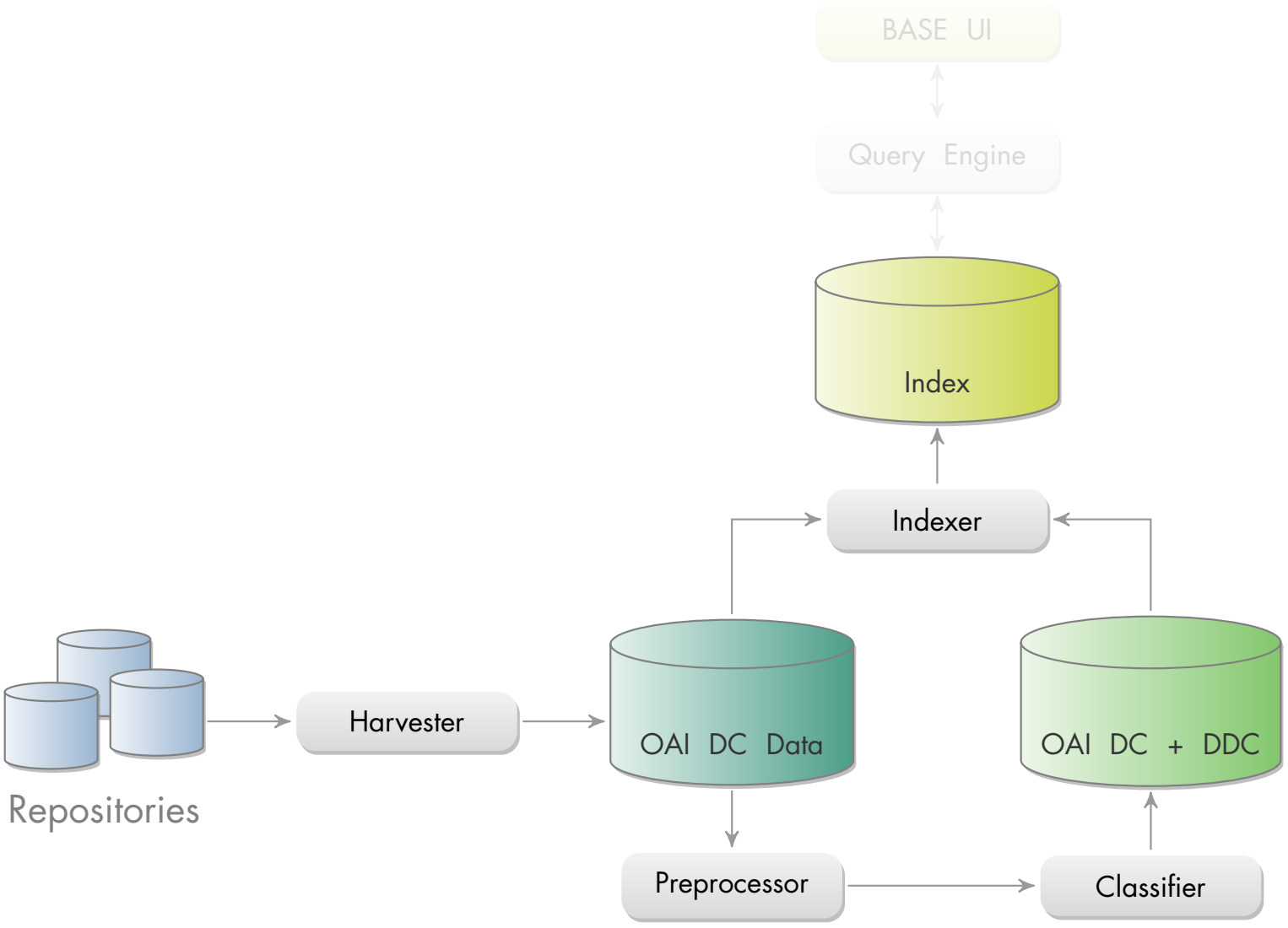
Metadatenanreicherung in BASE



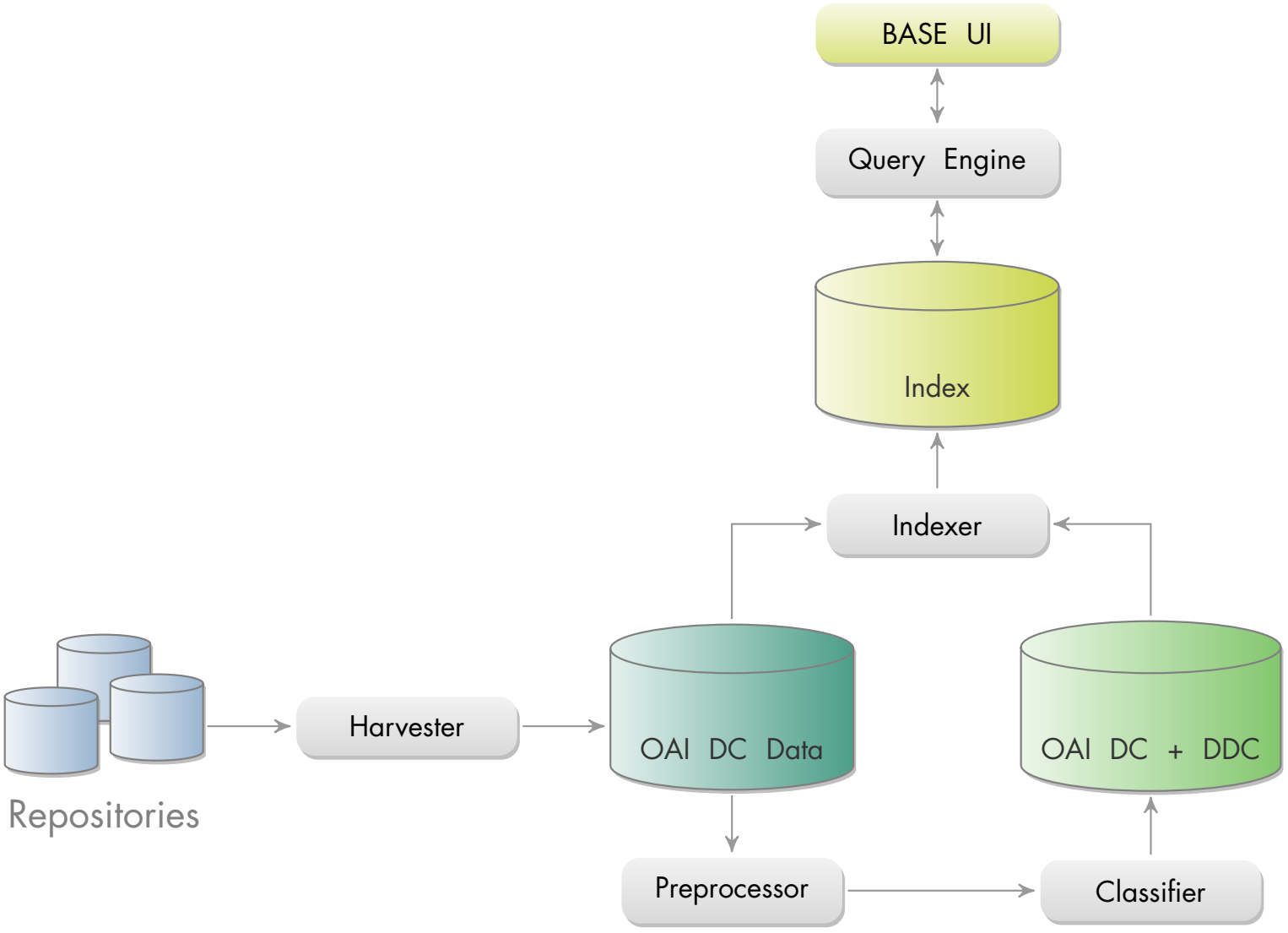
Metadatenanreicherung in BASE



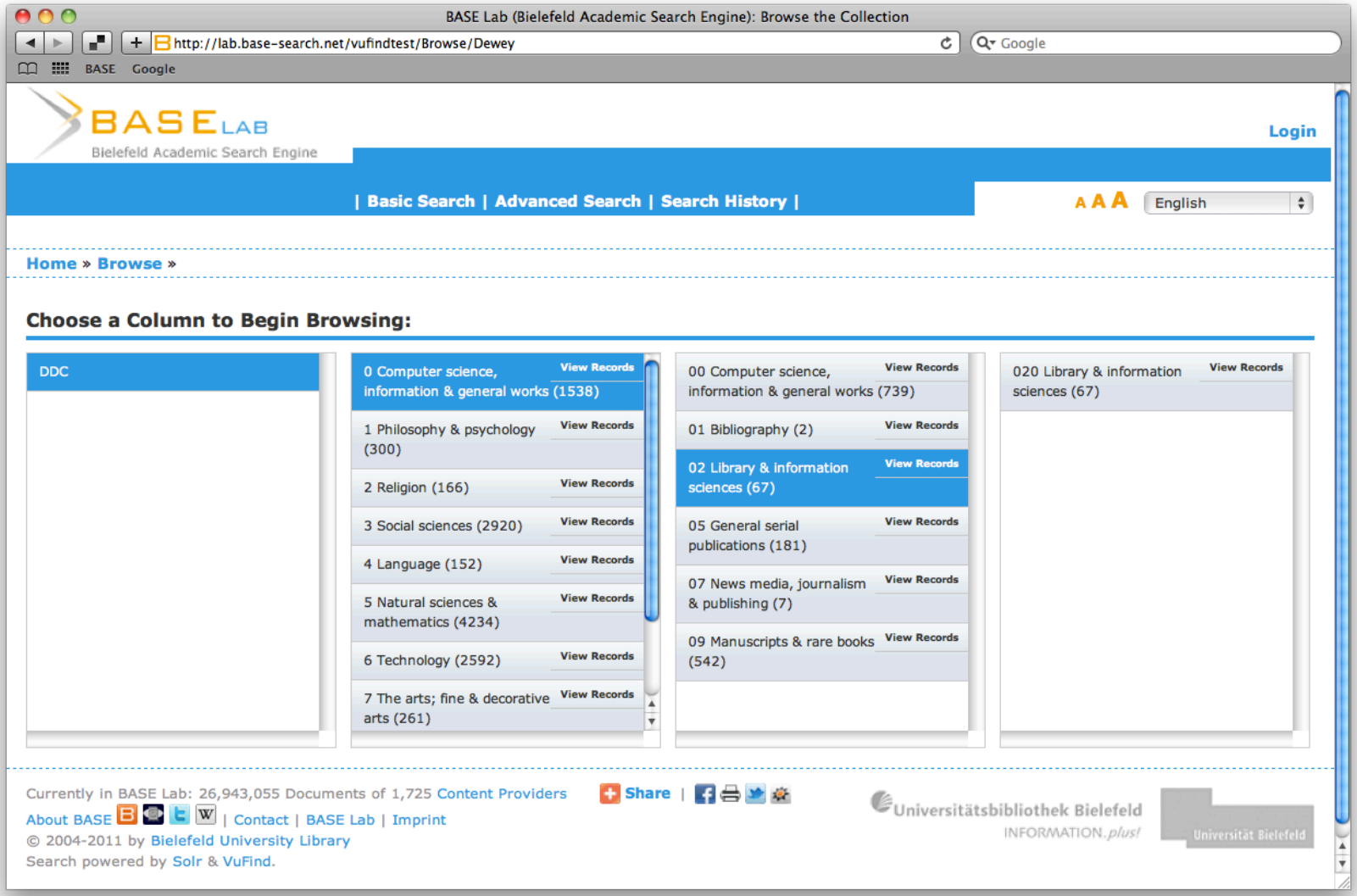
Metadatenanreicherung in BASE



Metadatenanreicherung in BASE



BASE Browsing



▸ Browsing aufrufen

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

Ergebnisse

Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

Intellektuell klassifizierte Dokumente:	443.249
Automatisch klassifizierte Dokumente:	522.871
DDC-klassifiziert total:	966.120
DDC-klassifiziert in Prozent:	3,08 %
Steigerung durch Automatisierung:	117,96 %

Ergebnisse

Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

Intellektuell klassifizierte Dokumente:	443.249
Automatisch klassifizierte Dokumente:	522.871
DDC-klassifiziert total:	966.120
DDC-klassifiziert in Prozent:	3,08 %
Steigerung durch Automatisierung:	117,96 %

Ergebnisse

Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

Intellektuell klassifizierte Dokumente:	443.249
Automatisch klassifizierte Dokumente:	522.871
DDC-klassifiziert total:	966.120
DDC-klassifiziert in Prozent:	3,08 %
Steigerung durch Automatisierung:	117,96 %

Ergebnisse

Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

Intellektuell klassifizierte Dokumente:	443.249
Automatisch klassifizierte Dokumente:	522.871
DDC-klassifiziert total:	966.120
DDC-klassifiziert in Prozent:	3,08 %
Steigerung durch Automatisierung:	117,96 %

Ergebnisse

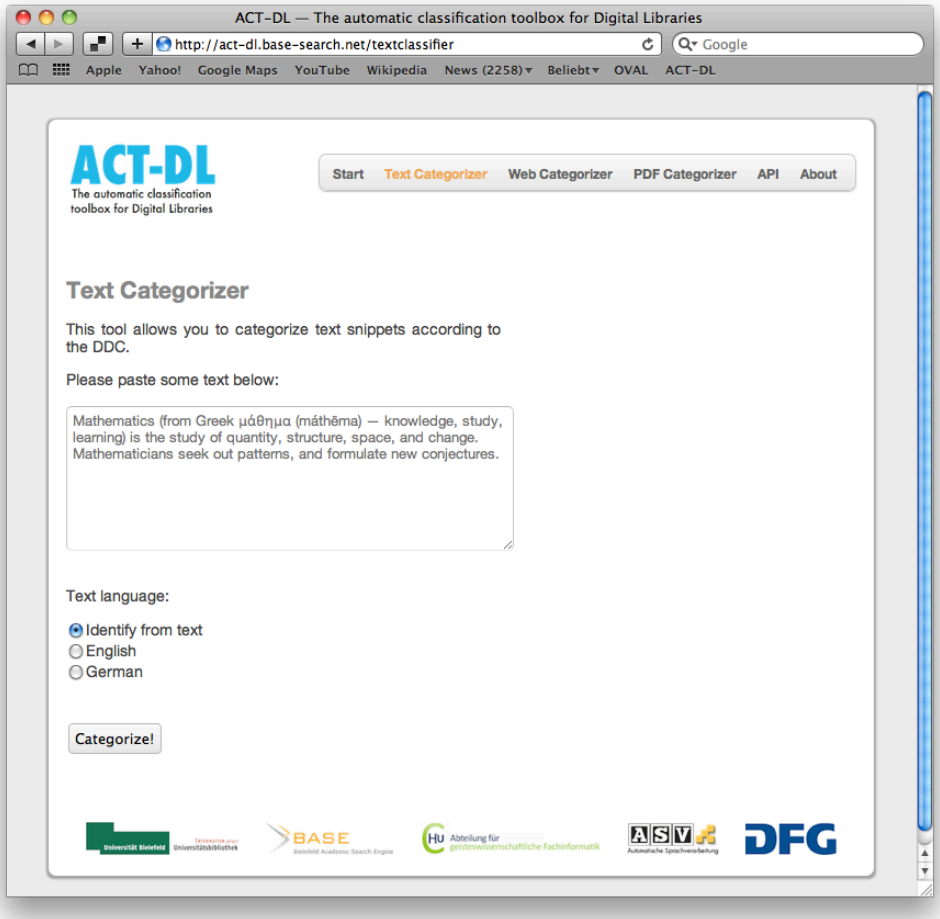
Produktive Klassifikation in BASE

Aktuelle Zahlen aus BASE:

Intellektuell klassifizierte Dokumente:	443.249
Automatisch klassifizierte Dokumente:	522.871
DDC-klassifiziert total:	966.120
DDC-klassifiziert in Prozent:	3,08 %
Steigerung durch Automatisierung:	117,96 %

ACT-DL

The Automatic Classification Toolbox for Digital Libraries



- Text Categorizer
- Web Categorizer
- PDF Categorizer

▸ Seite besuchen

Universität Bielefeld
Universitätsbibliothek
BASE Bielefeld Academic Search Engine
DFG

Nachnutzung: Klassifikator

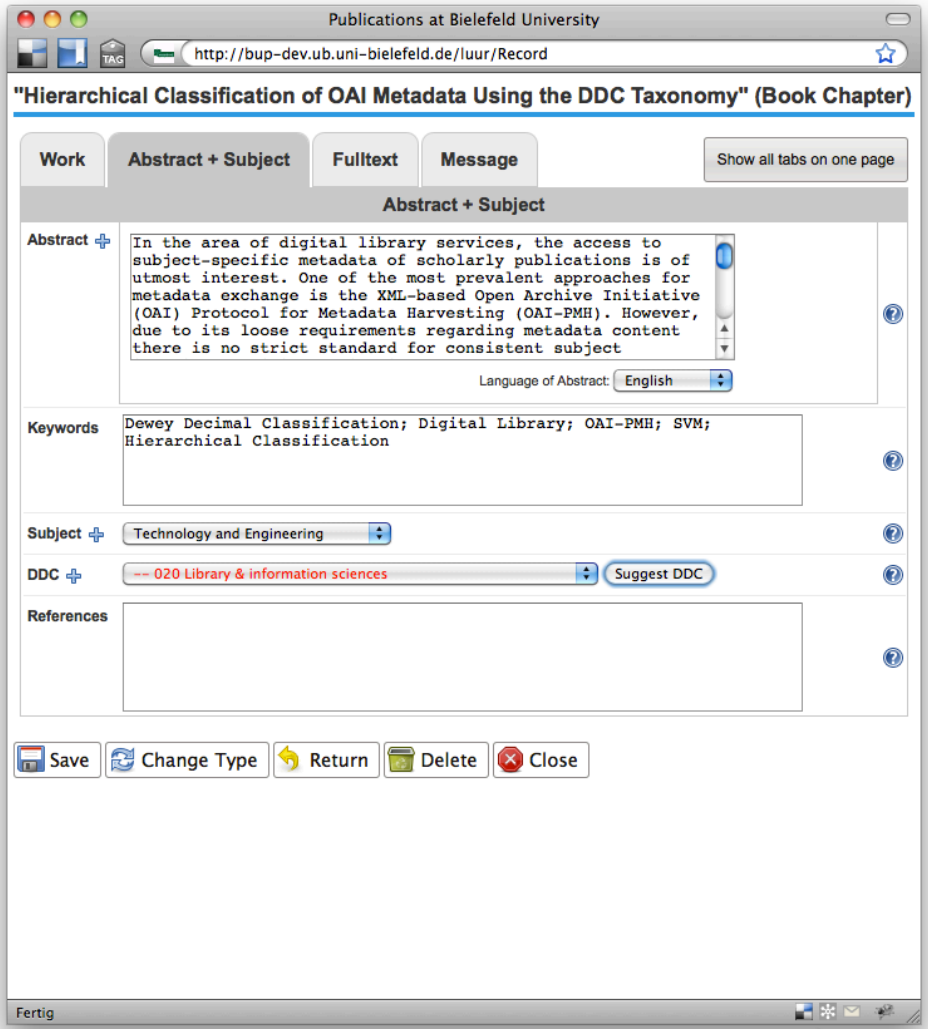
API

```
<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: Klassifikator

Vorschlagsystem für die Metadatenerfassung im Bielefelder Repository PUB



Universität Bielefeld

INFORMATION.plus! Universitätsbibliothek

BASE Bielefeld Academic Search Engine

DFG

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 Automatische Klassifikation

3 Ergebnisse

4 **Zusammenfassung**

Erkenntnisse

- Schwierigkeiten
  - Akquise von Trainingsdaten
  - Ab DDC Ebene 3: Abdeckung problematisch
- Erfolge
  - Grobklassifikation (1. und 2. Ebene) gut automatisierbar
  - automatische Vergabe von DNB-Sachgruppen (DINI-Empfehlung) auf jeden Fall erreichbar
  - semi-automatische Verfahren (Vorschlagssysteme) umsetzbar
- Ausblick
  - Verbesserung des Klassifikators: Erprobung anderer Algorithmen, interaktives Lernen durch intellektuelle Korrektur
  - Mehrfachklassifikation
  - Erforschung neuer Zielklassifikationen

Vielen Dank für die Aufmerksamkeit!

Universität Bielefeld | Universitätsbibliothek
Universitätsstr. 25
D-33615 Bielefeld

☎ +49 521 106-2546
✉ Mathias.Loesch@uni-bielefeld.de

http://base-search.net/
http://act-dl.base-search.net/

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

Literatur

Lösch, M., U. Waltinger, W. Horstmann, and A. Mehler (2011). Building a DDC-annotated corpus from OAI metadata. *Journal of Digital Information* *12*(2).

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.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys* *34*(1), 1–47.

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.