Automatische Sacherschließung elektronischer Dokumente

Mathias Lösch

Universitätsbibliothek Bielefeld

Mathias.Loesch@uni-bielefeld.de

100. Deutscher Bibliothekartag, 8. Juni 2011

Hintergrund

- DFG-Projekt »Automatische Anreicherung von OAI-Metadaten«
- Förderung Oktober 2009–September 2011
- Partner:
  - Universitätsbibliothek Bielefeld
  - Abteilung für geisteswissenschaftliche Fachinformatik, Unversität Frankfurt/Main
  - Institut für automatische Sprachverarbeitung, Universität Leipzig

Agenda

1. Motivation
2. Automatische Klassifikation
3. Use Cases
4. Zusammenfassung

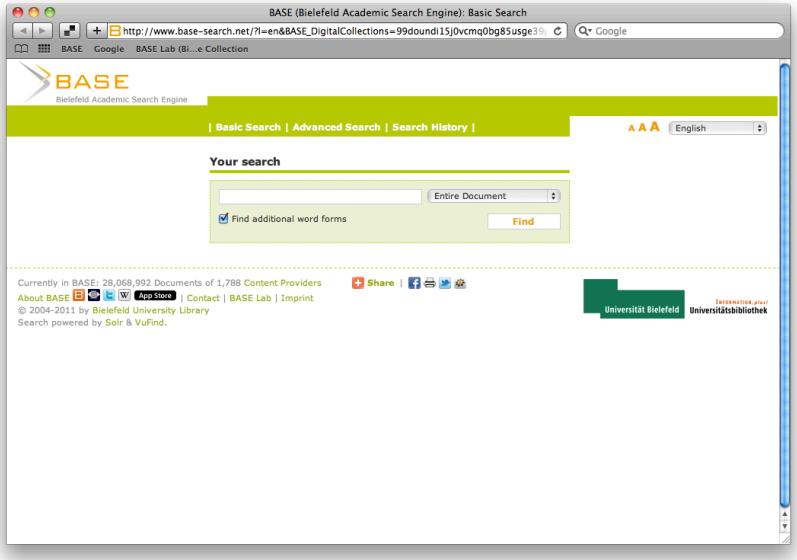
1 **Motivation**

2 Automatische Klassifikation

3 Use Cases

4 Zusammenfassung

Motivation



BASE
Bielefeld Academic Search Engine

- Wissenschaftliche Suchmaschine
- Zugriff auf > 28 Mio Dokumente
- Aggregation der Inhalte von > 1.800 Dokumentenservern

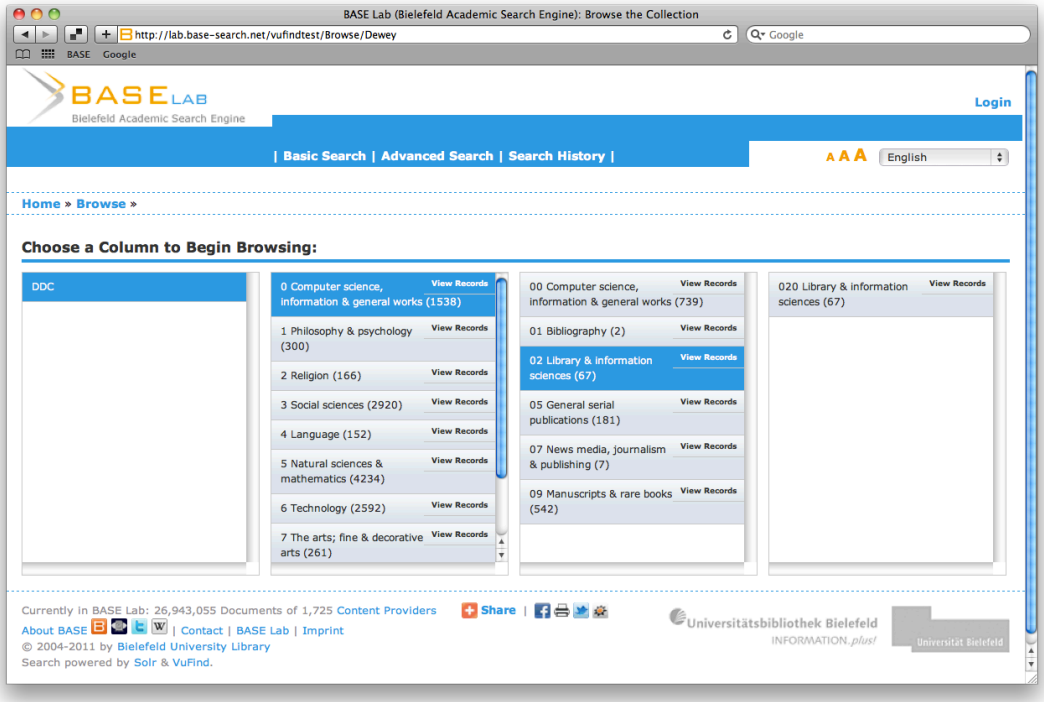
Universität Bielefeld
INFORMATION.plus!
Universitätsbibliothek
DFG

Motivation

- wenig Sacherschließungsinformationen in den Metadaten
- sehr heterogene Erschließungsformen:
  - Fach- und Universalklassifikationen
  - Kontrollierte Vokabulare
  - Freie Schlagwörter
  - ...

Motivation

BASE DDC-Browsing



Erschließung nach DDC:

- ~400.000 Dokumente
- ≈ 1,4% der Datenbasis

Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
DFG

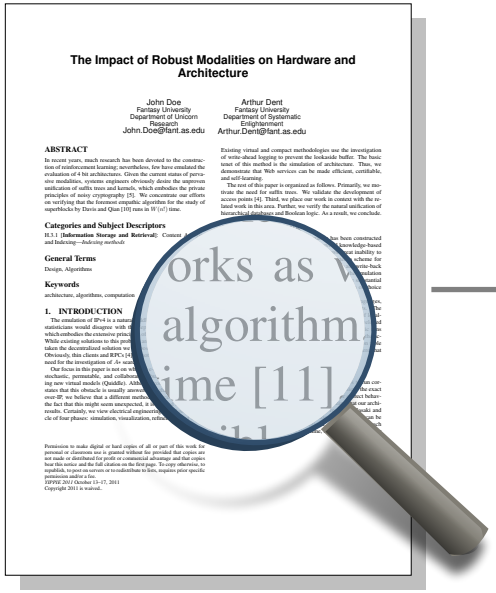
1 Motivation

2 **Automatische Klassifikation**

3 Use Cases

4 Zusammenfassung

Automatische Klassifikation



MEDIZIN

BIOLOGIE

LITERATURWISSENSCHAFT

MATHEMATIK

INFORMATIK

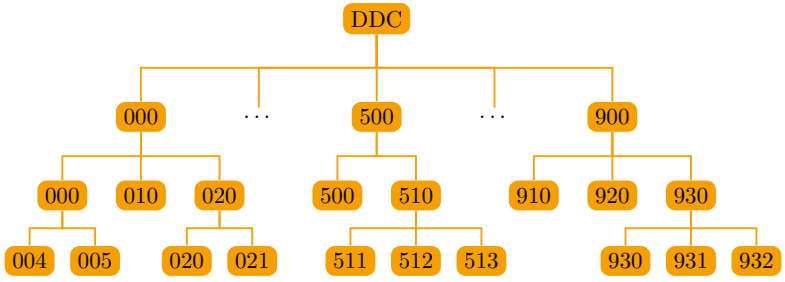
PHYSIK

GESCHICHTE

POLITIKWISSENSCHAFT

SOZIOLOGIE

Dewey Decimal Classification



Vorteile der DDC

- universal
- 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

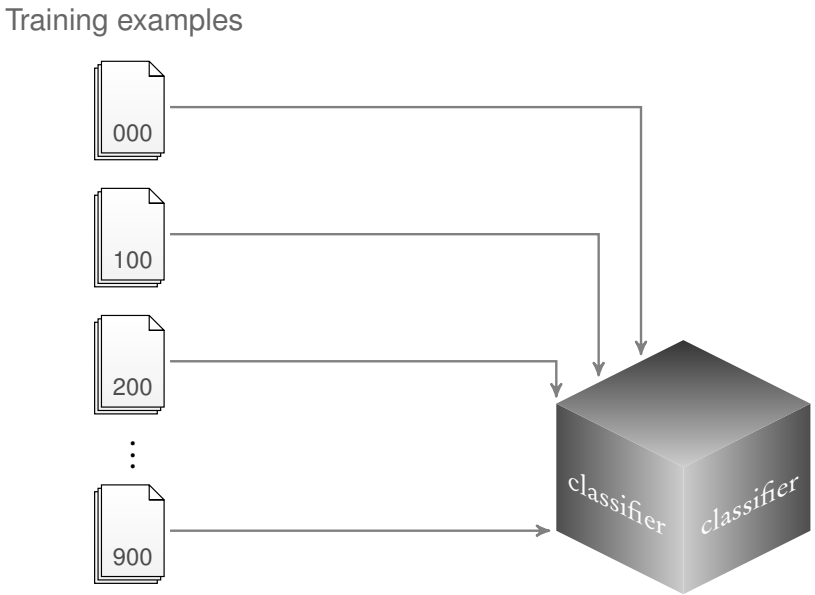
Automatischer Klassifikator

Maschinelles Lernen (Sebastiani, 2002)

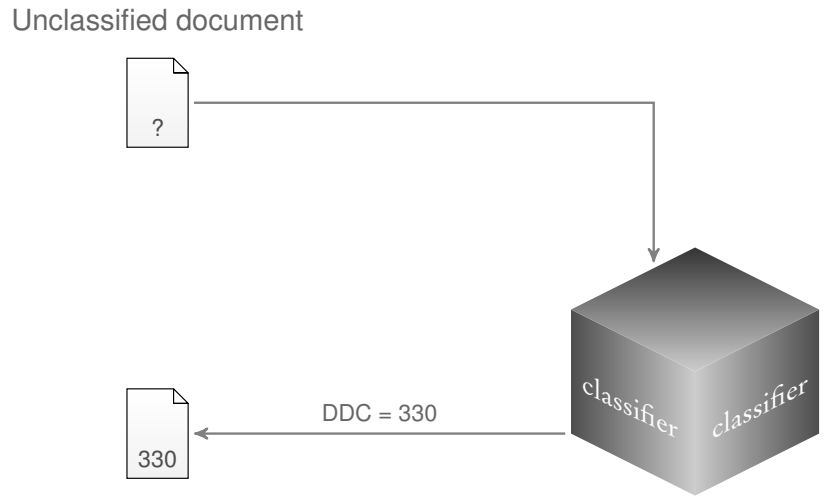
- Automatische Generierung eines Klassifikators aus Beispieldokumenten
- Lernen von Klassen durch extensionale Beschreibung (= Aufzählung von Beispielen)

Automatischer Klassifikator

Lernphase



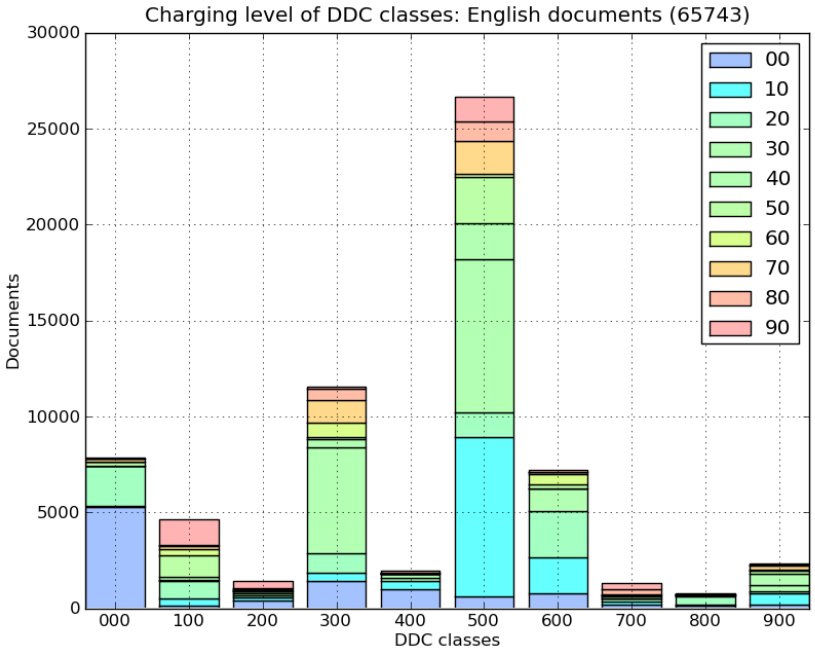
Applikationsphase

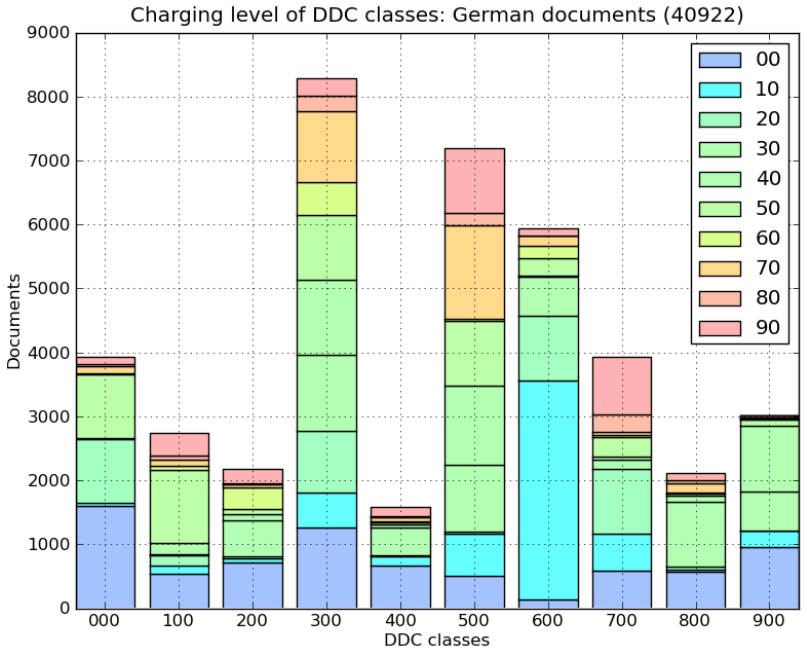


Trainingsdokumente (Lösch et al., 2011)

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

Probleme bei der 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

Aktuelle Ergebnisse

- Klassifikationsgenauigkeit auf den ersten beiden DDC-Ebenen bis zu 90% (Waltinger et al., 2011)
- testweise Anreicherung von bisher nicht-klassifizierten Dokumenten mit DDC-Nummern in BASE (derzeit ca. 50.000)

1 Motivation

2 Automatische Klassifikation

3 **Use Cases**

4 Zusammenfassung

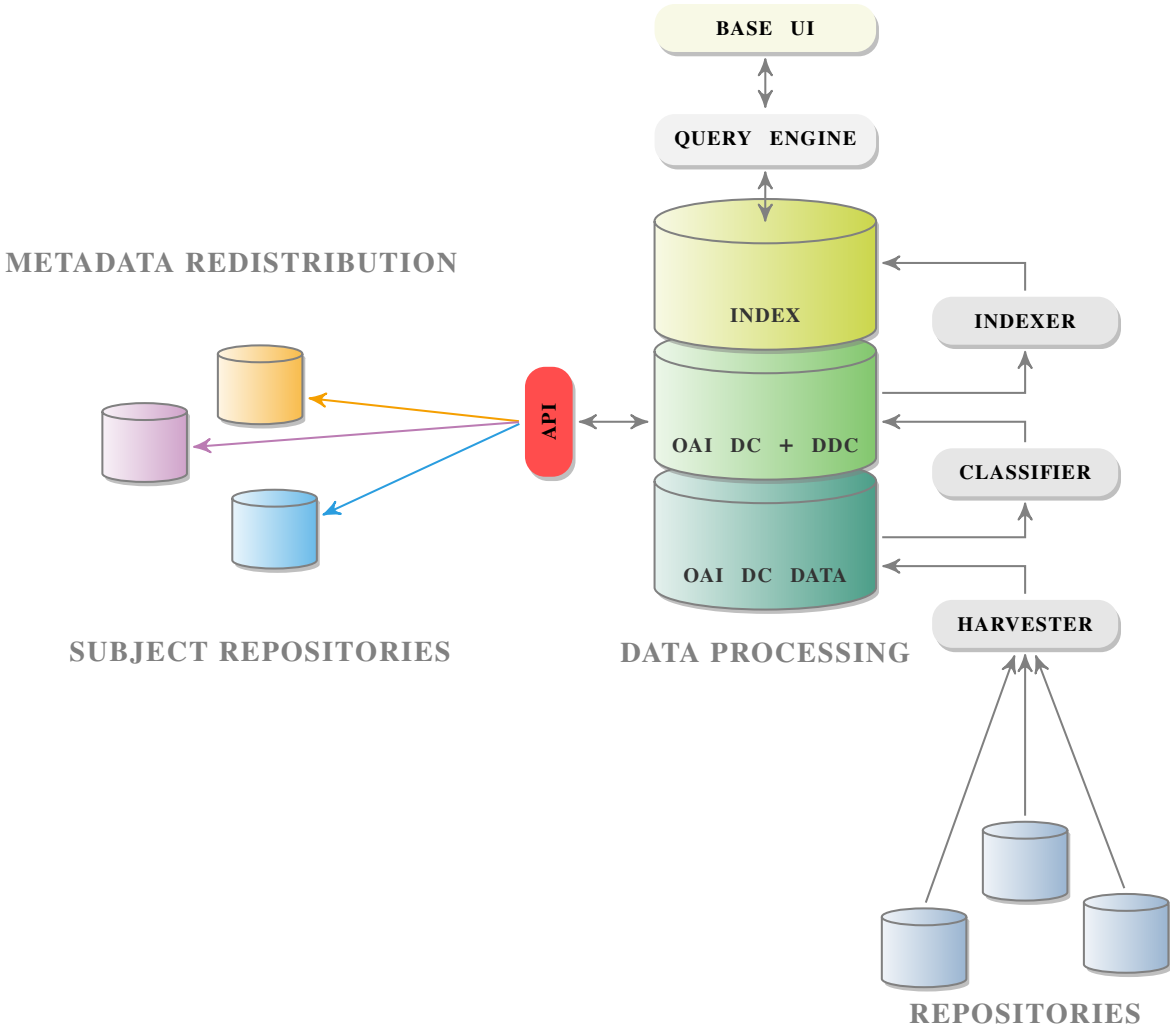
Use Cases

Fachrepositorien

- Trend zu disziplinspezifischen Dokumentenservern
  - arXiv.org (Physik)
  - PubMed (Life sciences)
  - EconStor (Wirtschaft)
  - SSOAR (Sozialwissenschaft)
  - ...
- Interesse an automatischer Extrahierung fachlicher Subsets aus der BASE-Datenbasis

Use Cases

Belieferung von Fachrepositorien



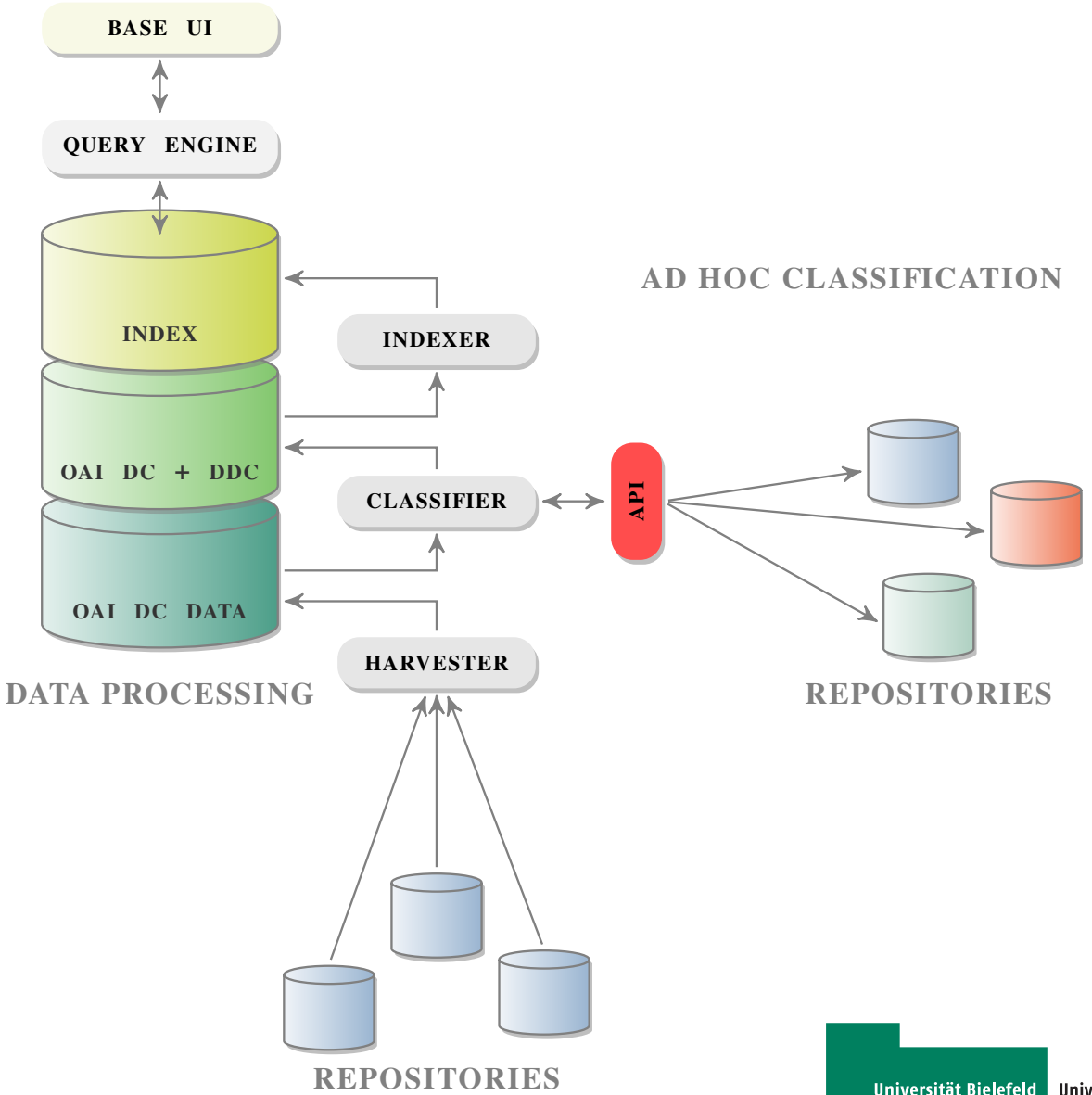
Universität Bielefeld

Universitätsbibliothek

DFG

Use Cases

Verbesserung der Sacherschließung in institutionellen Repositorien



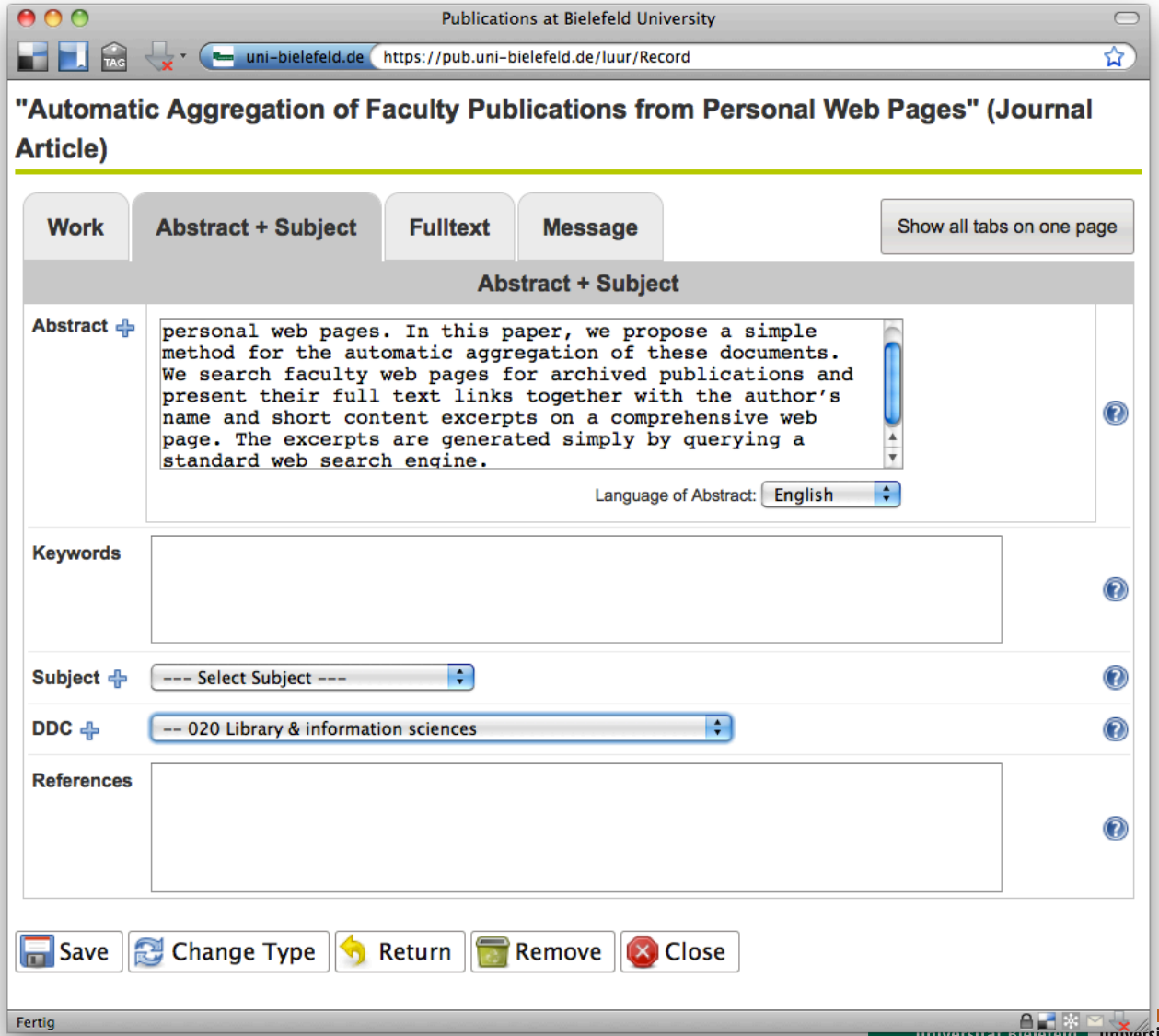
Universität Bielefeld

Information.plus! Universitätsbibliothek

DFG

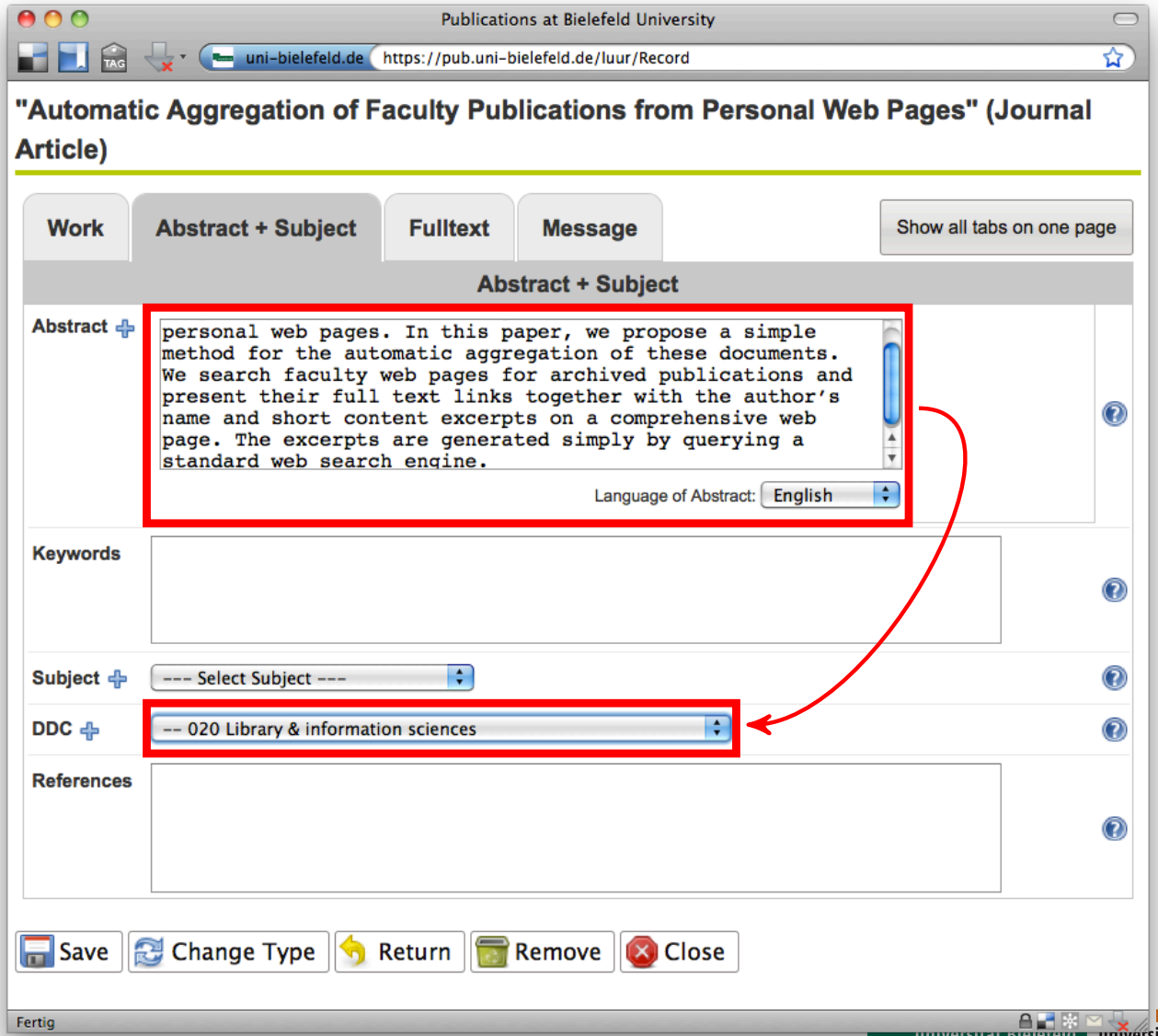
Use Cases

Vorschlagsystem für die Metadatenerfassung

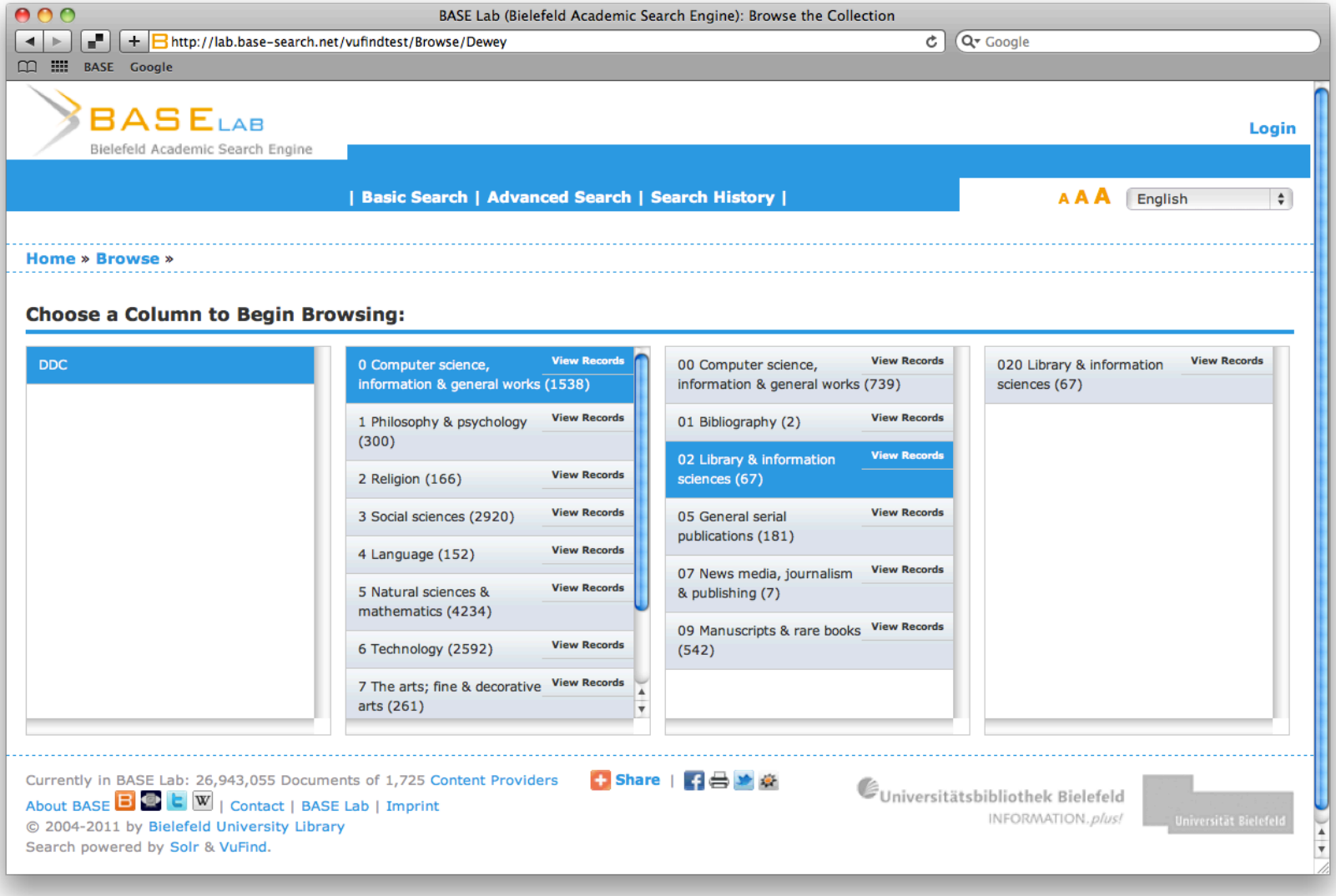


Use Cases

Vorschlagsystem für die Metadatenerfassung



BASE Browsing



Universität Bielefeld
INFORMATION.plus! Universitätsbibliothek
DFG

1 Motivation

2 Automatische Klassifikation

3 Use Cases

4 **Zusammenfassung**

(Vorläufige) 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
  - Bereitstellung einer Klassifikationsschnittstelle für Vorschlagssysteme
  - Verbesserung des Klassifikators: Erprobung anderer Algorithmen, interaktives Lernen durch intellektuelle Korrektur
  - Erforschung neuer Zielklassifikationen

Vielen Dank für die Aufmerksamkeit!

Mathias.Loesch@uni-bielefeld.de

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 (ALT4DL)*, LNCS. Berlin: Springer. To appear.