FIS Bildung Tagung 2010

DIPF
Bildungsforschung
und Bildungsinformation

Kurzpräsentationen des Werkstattgesprächs zum Thema „Unterstützung der Dokumentenerschließung durch automatisierte Verfahren“ am 04.05.10:

- Verena Wissel: Automatisierte Verfahren im Bereich der Sach- u Formalerschließung (**DIPF**)
- Christa Schöning-Walter: Erfahrungsbericht der **DNB** im Rahmen des Projekts PETRUS
- Manfred Faden/Thomas Groß: Erfahrungsbericht der **ZBW** zum MindServer Categorizer
- Monika Zimmer: Erfahrungsbericht der **GESIS** zum MindServer Categorizer in SOLIS
- Michael Gerards: Erfahrungsbericht des **ZPID** zur Arbeit mit AUTINDEX

DIPF
Bildungsforschung
und Bildungsinformation

Automatisierte Verfahren

FIS Bildung Tagung im Mai 2010

DIPF
Bildungsforschung
und Bildungsinformation

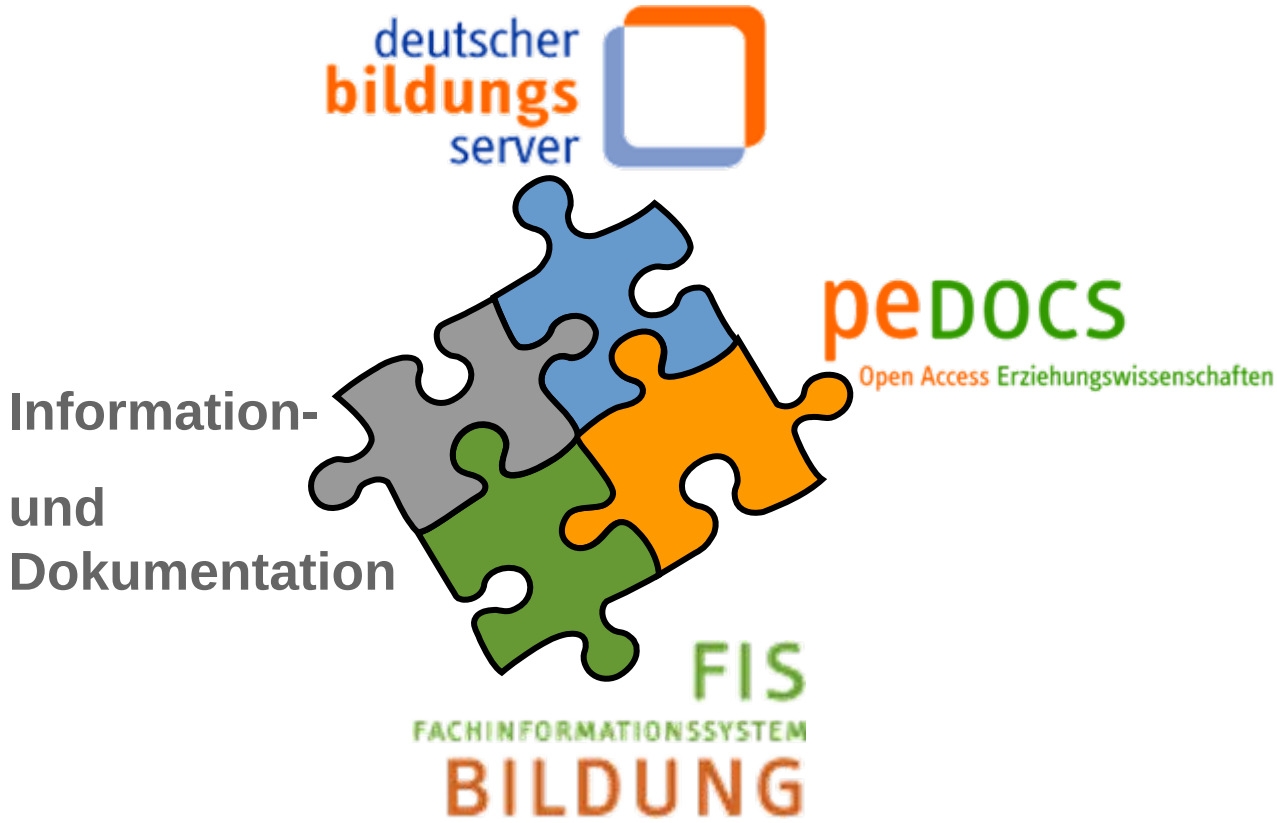
Übersicht

- Ziele und Vorgehen der DIPF-AG
- Bestands- und Bedarfsanalyse der einzelnen Teams
- Verfahren zur Automatischen Indexierung
- Zusammenhang mit Suchmaschinentechnologie
- Marktübersicht
- Ausgewählte Unternehmen
- Weitere Schritte
- Vorläufiges Fazit

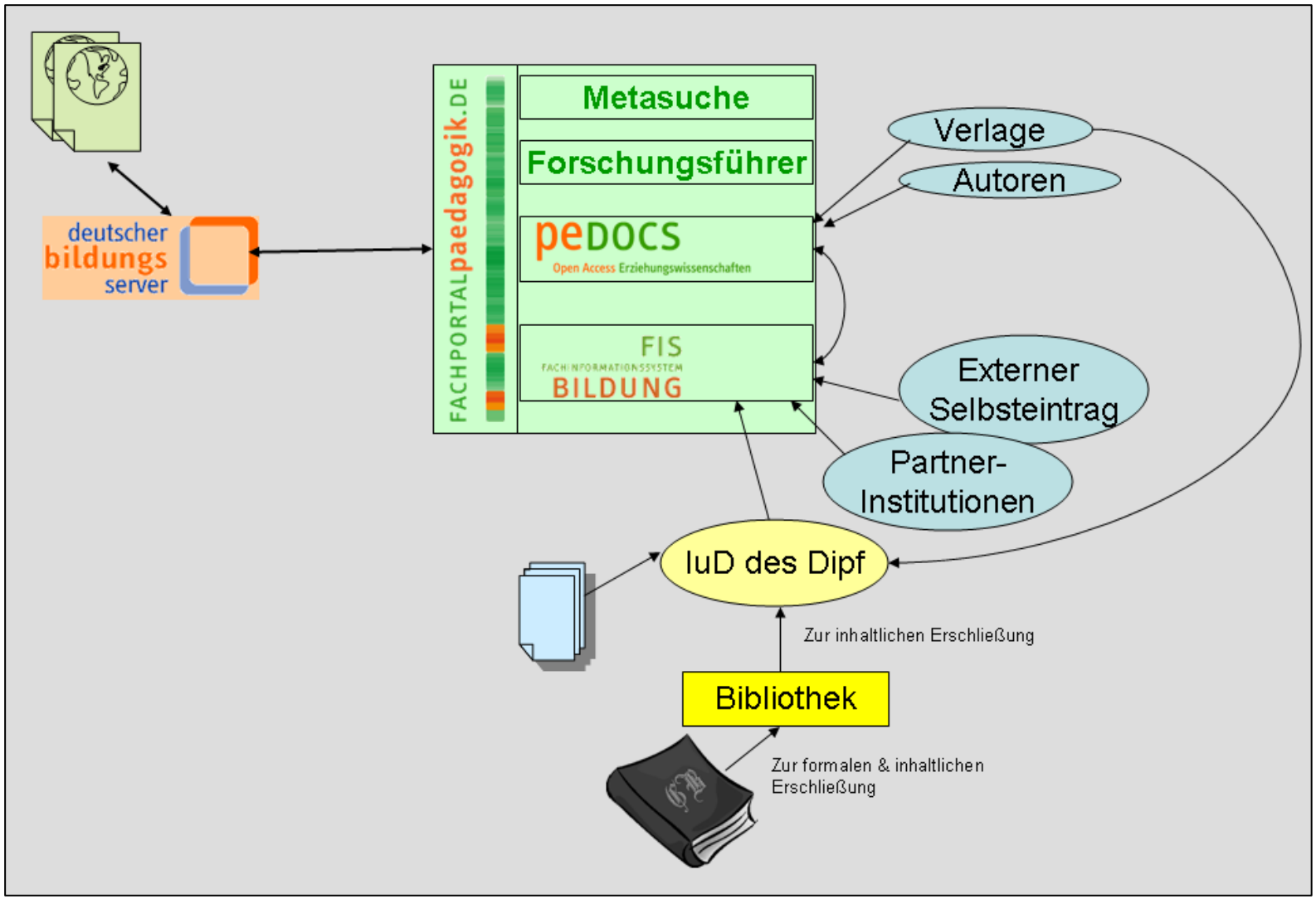
Ziele und Vorgehen der AG

- Ziel: Automatisierung einzelner Arbeitsabläufe im Workflow des DIPF zur Beschleunigung und Erleichterung einzelner Arbeitsschritte und zur Verbesserung des Retrievals
- Vorgehen bis Ende 2010
  - Bestands- und Bedarfsanalyse am DIPF für folgende Dienste: FIS, Pedocs, IuD und DBS ✔
  - Sichtung von Anbietern und Produkten am Markt ✔
  - Sammlung von Erfahrungsberichten anderer Institutionen ✔
  - Auswahl geeigneter Anbieter
  - Testung und Evaluierung von Verfahren
  - nach 2010: evtl. Implementierung eines Verfahrens

Zusammensetzung der AG



Bestandsanalyse I



Bestandsanalyse II

- Unterschiedliche Datenquellen und -Formate
- Übernahme der Daten überwiegend per copy & paste
- Keine automatische Indexierung
- Catalogue Enrichment
  - In FIS 30% aller Monografien mit CE-Link (davon 70% zu Inhaltsverzeichnissen)
  - Pedocs gibt CE-Dateien teils an FIS weiter
  - IuD hat nur Links zu CE-Dateien in Datenbank, die über OPAC nicht suchbar sind

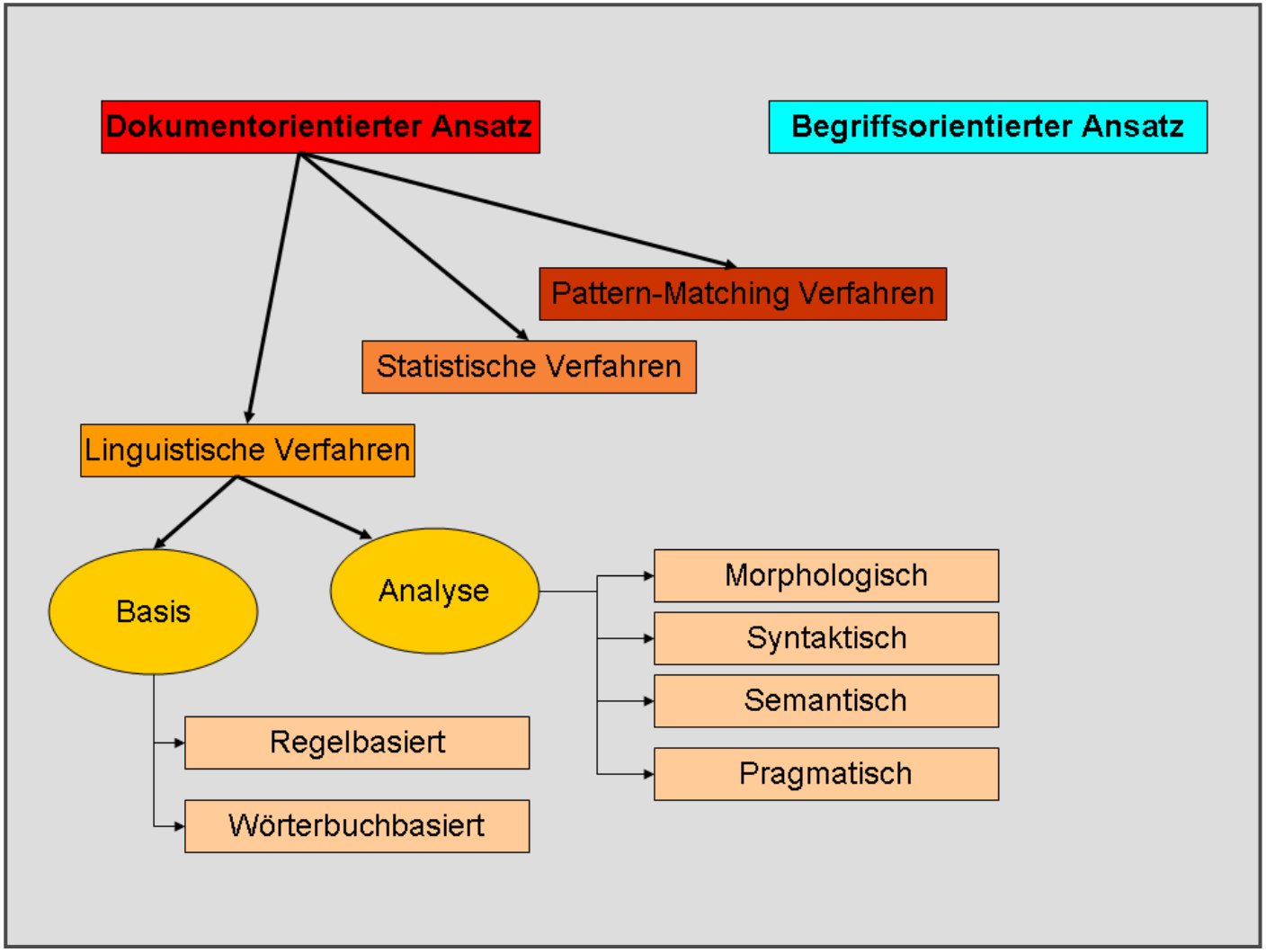
DIPF
Bildungsforschung
und Bildungsinformation

Bedarfsanalyse

- **Datenerfassung von**
  - Katalogdaten / Metadaten
  - Catalogue Enrichment Daten (Inhaltsverzeichnisse)
  - Volltexten (Monografien, Sammelwerksbeiträge, Aufsätze)
  - Hochschul-Volltextserver (Abtracts und Schlagworte)
  - Retrodigitalisierte Dokumente evtl. mit OCR-Fehlern
- **Datenerschließung**
  - Voll- oder Semi-Automatische Erschließung
  - In unterschiedlichen Sprachen & Formaten
  - Schlagworterkennung trotz OCR-Fehler
  - Schlagwortgenerierung aus Systematikstellen Klassifikation
  - Möglichkeit der Einbindung und Bearbeitung des (eigenen) Vokabulars
  - Abgleich zugelieferter Deskriptoren mit FIS Schlagwörtern
  - Kategorisierung: Automatische Zuordnung von Dokumenten zu pädagogischen Teildisziplinen

DIPF
Bildungsforschung
und Bildungsinformation

Verfahren zur Automatischen Indexierung



DIPF
Bildungsforschung
und Bildungsinformation

Zusammenhang mit Suchmaschinentechnologie

- Lucene ist implementiert und wird aktuell angepasst
- Beim DBS derzeit inhaltliches Ranking und Gewichtung
- In FIS Literaturdatenbank bisher nur Einfluss auf Performanz
- Für die Komplexität der Deutschen Sprache ist eine „Vorab-Sprachverarbeitung“ sinnvoll
- Automatische Indexierung zeigt erst beim Retrieval (Sortierung) Wirkung
- Effekt: Erhöhung Recall
- ABER: zur Anpassung der Precision ggf. auch Verbesserung der Querys nötig, z.B. durch Einbeziehung eines kontrollierten Vokabulars (FIS Thesaurus)
- Fragwürdig: was passiert mit Dokumenten, die noch zu Zeiten vor der automatischen Indexierung erschlossen wurden?

Marktübersicht

temis

Giai

intelligent views

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

AGI INFORMATION MANAGEMENT CONSULTANTS

brainbot TECHNOLOGIES AG

ImageWare

DIGITAL COLLECTIONS

INTRAFIND

IBM

SER

EMPOLIS an attensity group company

inxight

picturesafe

living-e

SEARCH CLASSIFY NAVIGATE X SEARCH

textwise

iSquare

recommind software that understands

averbis medical language technology

Lingsoft LANGUAGE SOLUTIONS

TEX TEC SOFTWARE

LIBERO INTEGRIERTES BIBLIOTHEKS-MANAGEMENTSYSTEM

rapid-i REPORT THE FUTURE

Ausgewählte Unternehmen

AGI
INFORMATION
MANAGEMENT
CONSULTANTS

recommind
software that understands

SEARCH CLASSIFY NAVIGATE
SEARCH

temis

picturesafe

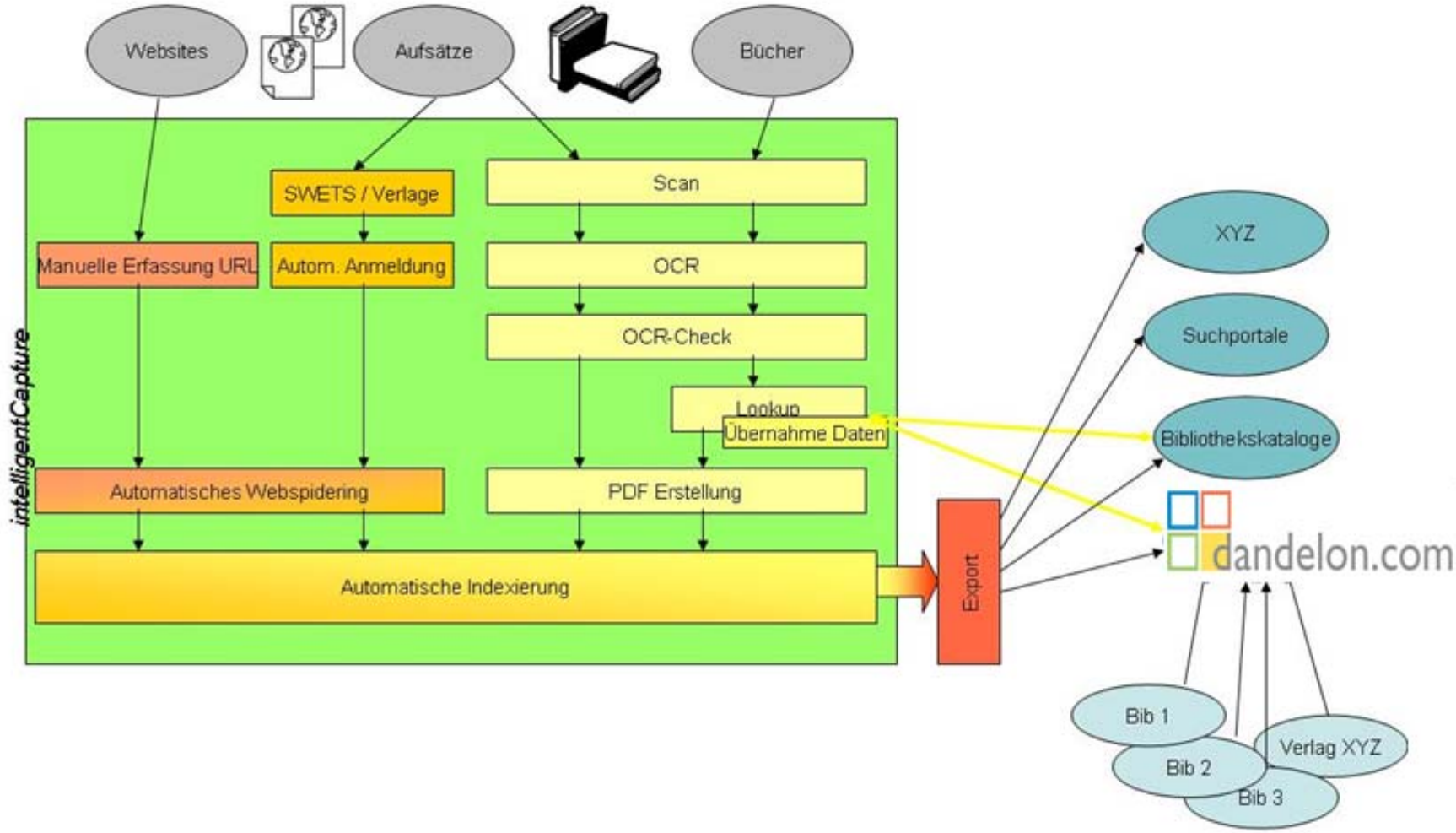
TEX
TEC
SOFTWARE

EMPOLIS
an otterstedt group company

IntelligentCapture von AGI

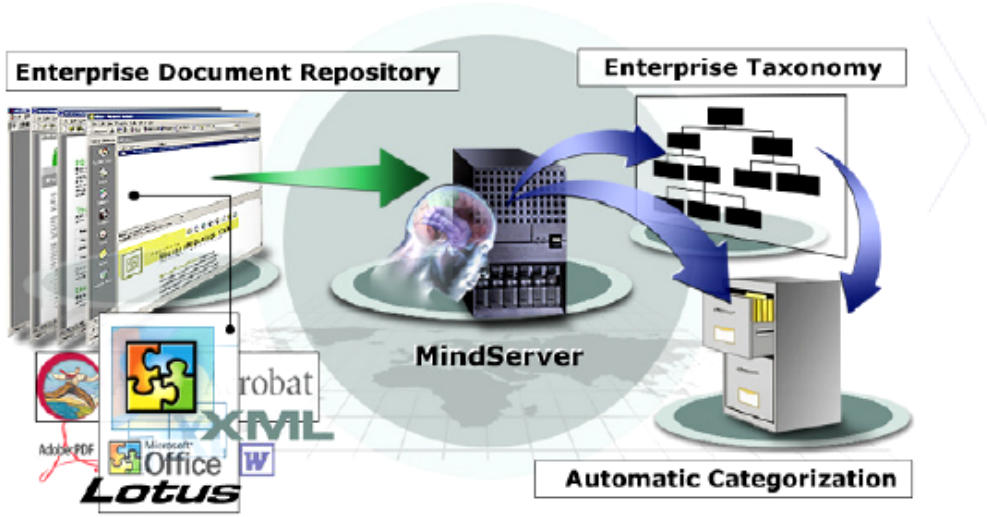
- Mobile Scan-Stationen & OCR-Software inklusive Fehlererkennung
- Linguistische & Statistische Sprachverarbeitung mit AUTINDEX
- Einbindung von Text Mining Lösungen durch Zusammenarbeit mit Temis
- Bedingte Erschließung von Einzelwerken aus Sammelbeiträgen mit Pen möglich; zur Automatisierung darf sich Layout einer Zeitschrift nicht verändern
- FineReader 9: neue Sprachen (Chinesisch, Thailändisch, usw.)
- Prototyp „Fetchpaper“: automatisierte Abholung von PDFs auf Internetseiten
- Verarbeitung mehrerer Texttypen in einem Dokument
- Keine Kategorisierung

IntelligentCapture von AGI

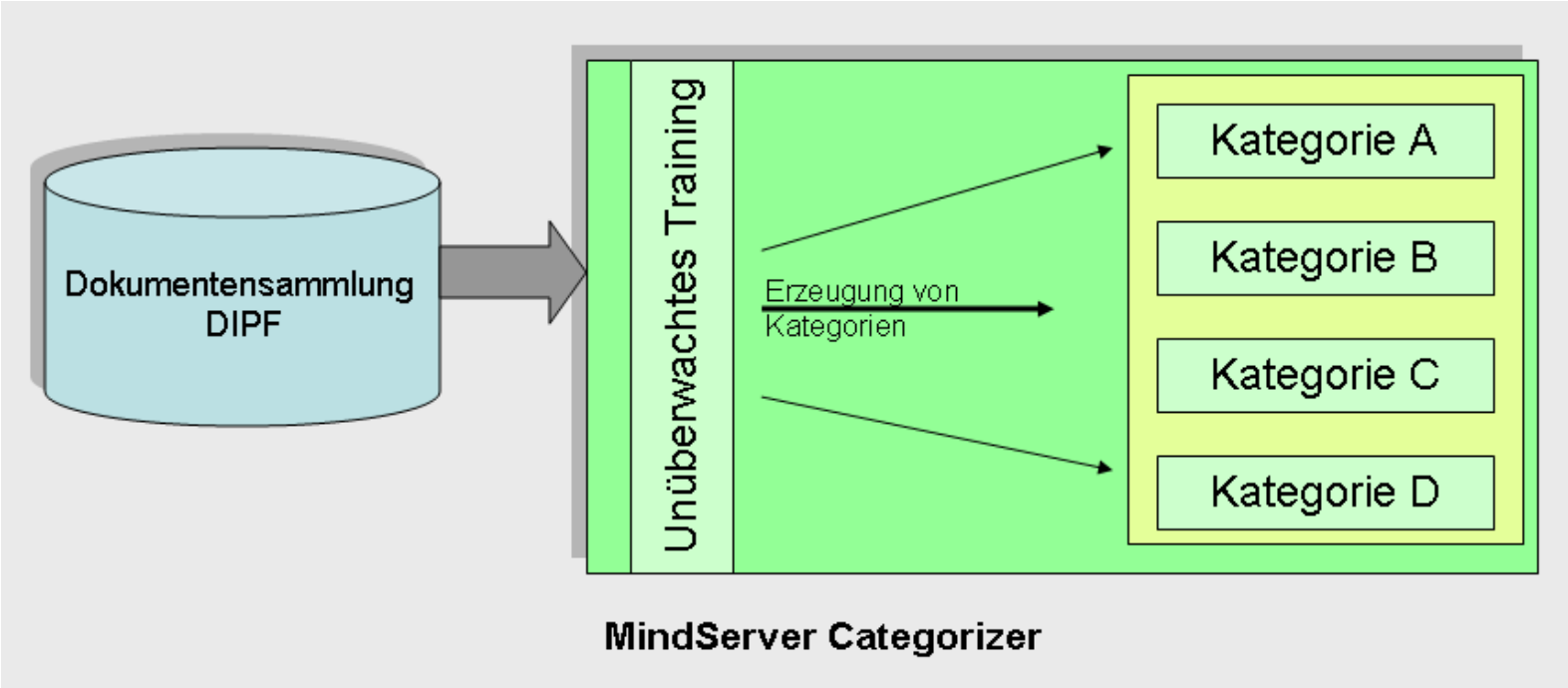


MindServer Categorizer von Recommind

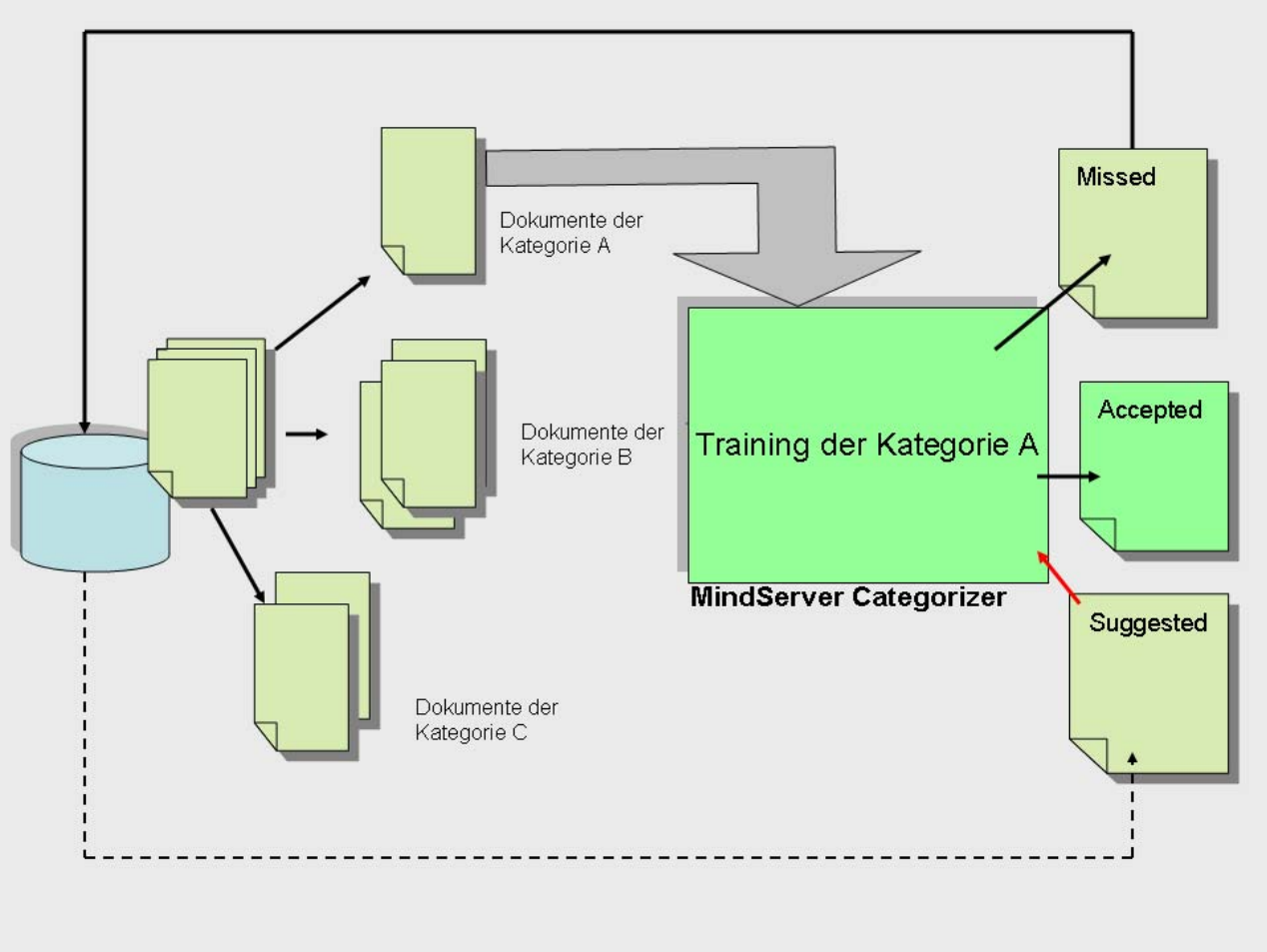
- Verarbeitung von 30 Sprachen und 350 Formaten
- Statistische Verarbeitung: *Probabilistic Latent Semantic Indexing* Algorithmus
- Linguistische Analyse und Wörterbuch-Einbindung sind optional
- Unüberwachtes oder Überwachtes Training zur Kategorisierung



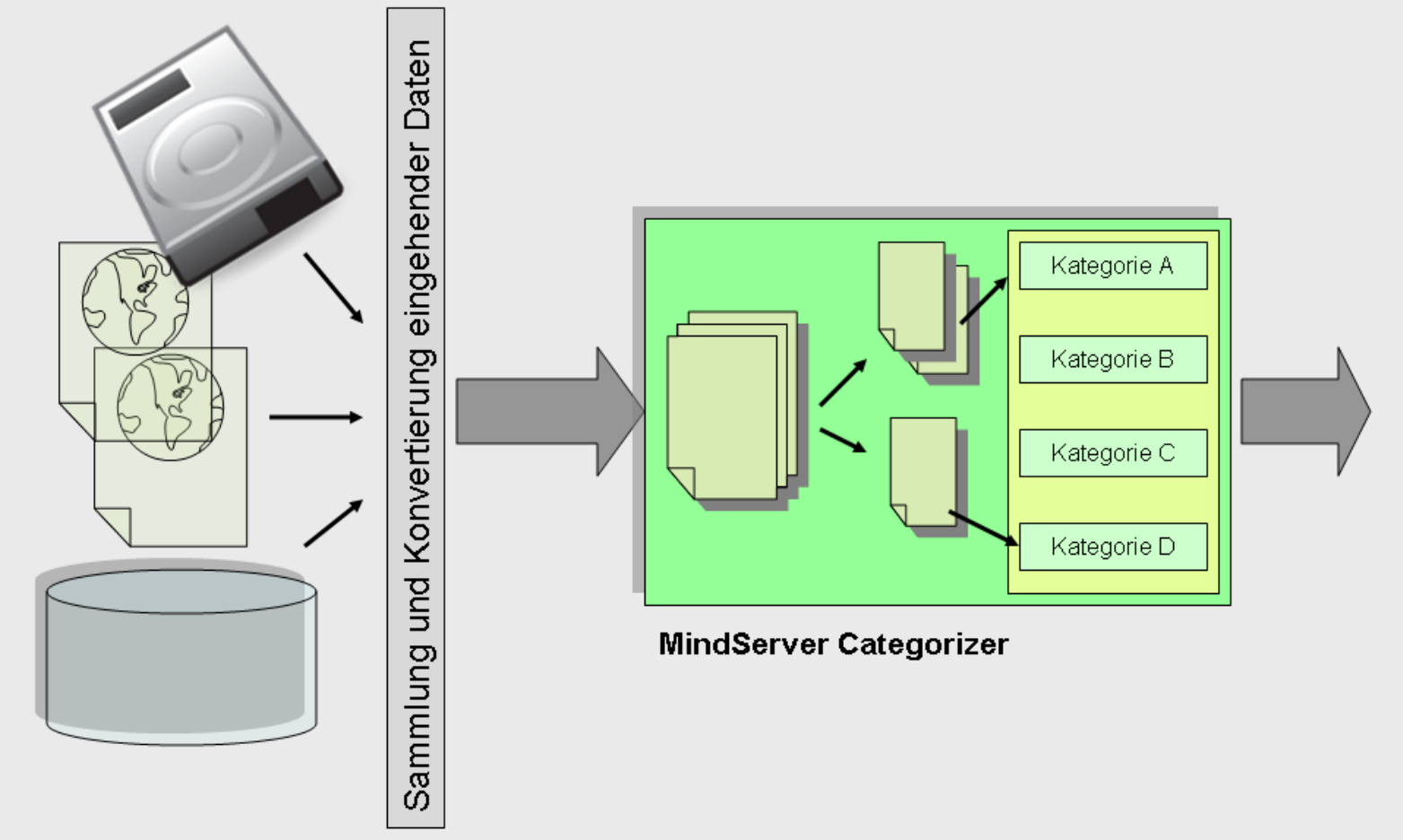
Unüberwachtes Lernen



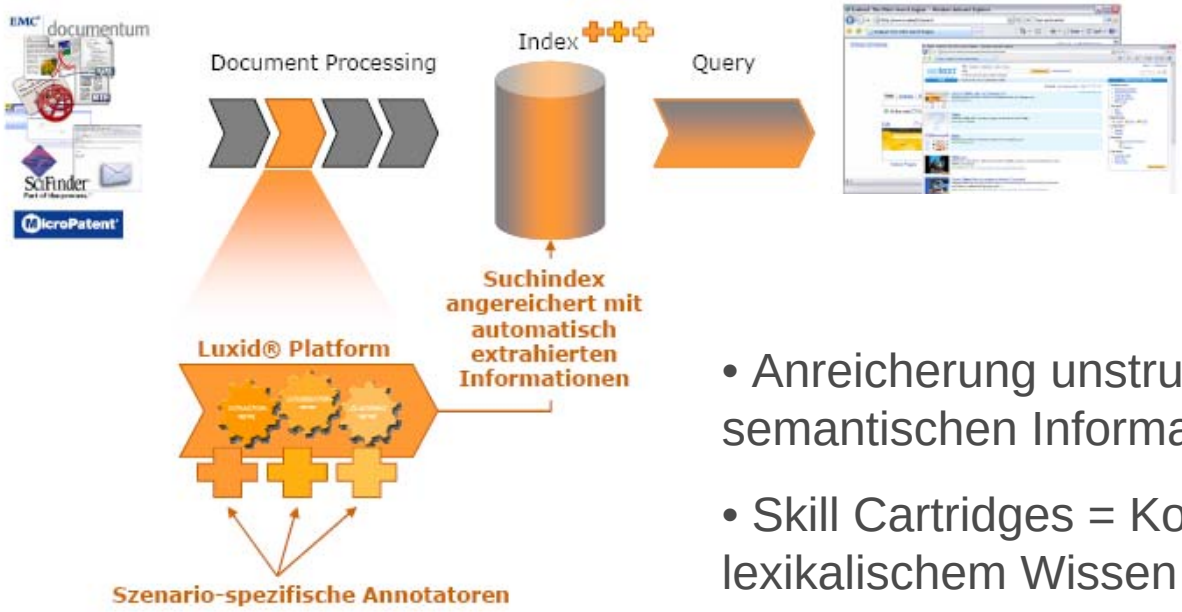
Überwachtes Lernen



Kategorisierung



Luxid von Temis

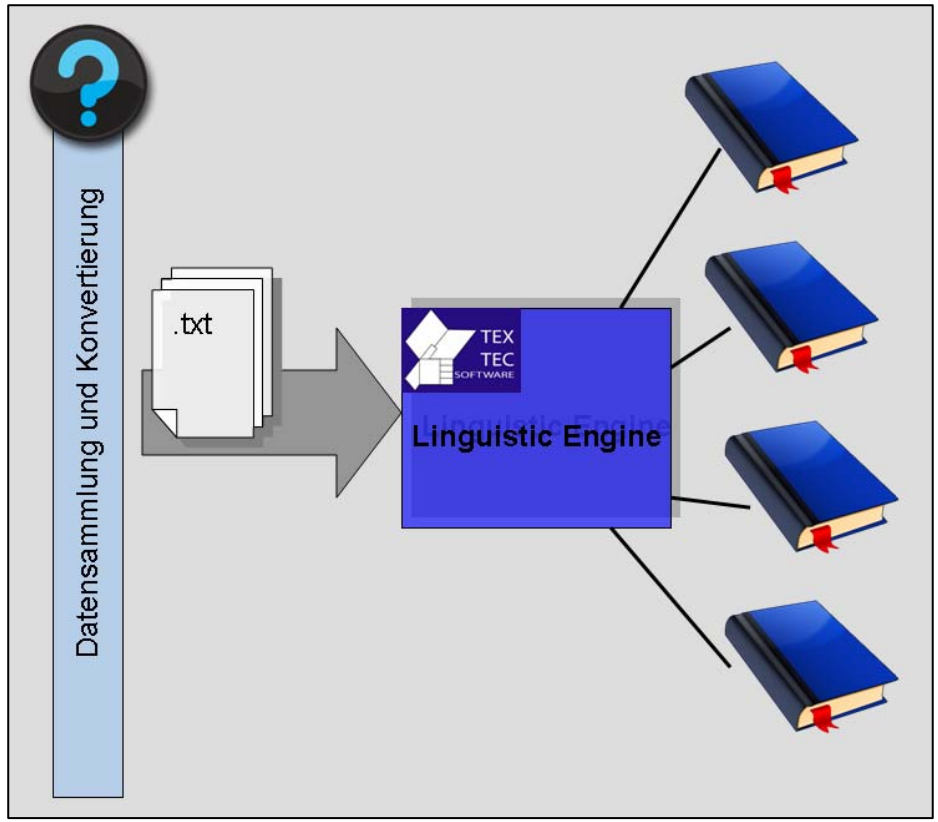


- Anreicherung unstrukturierter Texte mit semantischen Informationen
- Skill Cartridges = Kombination aus lexikalischem Wissen und Regeln
- Verschiedene Analyseebenen:

Basis-Sprachverarbeitung → Term Extraktion → Entitäten-Extraktion → Syntaktisches Parsing → **Fakten/Event Extraktion**

Weitkämper / Textec

- Verarbeitung von Textdateien
- Bis zu 100 Wörterbücher (Vollformen)
- Erstellung neuer Wörterbücher



Anbieter im Vergleich

DIPF
Bildungsforschung
und Bildungsinformation

	AGI	Recommind	Temis	Textec	Empolis	Picturesafe
Scan/OCR	✔	-	-	-	-	-
Erfasste Formate	Scan: txt, HTML, TIF; MARC, MAB, XML, ONIX, SWETS u.a.	350	txt, HTML, XML, Binärformat, kein Image-PDF!	-	300	-
Verarbeitetes Format	PDF	XML		Txt	300 (am Ursprungsort!)	XML
Konvertierung	✔	✔	✔	-	-	-
Verfahren zur Sprach-verarbeitung	linguistisch – statistisch	statistisch	linguistisch – statistisch	linguistisch	linguistisch – semantisch	statistisch
Kategorisierung	-	✔	✔	-	✔	✔
Semantische Technologien	Text Mining (Temis)	-	Taxonomien, semantische Netze	-	Ontologien, semantische Netze	-
Sprachen	4 Indexierung 150 OCR	30	20	8-9	32	beliebig
Metadaten-anreicherung	Semi-automatisch mit Scan-Stift	✔	✔	-	✔	✔
Erkennung von Entitäten	Temis	✔	✔	✔	✔	✔
Kosten In Euro	8.500 + 2-12.000/Jahr + Customizing	?	150.000 + 35.000/Jahr + 20% Wartung	15-35.000 + 16% Wartung	?	Kaufoption: 44.030 + 16-22% Wartung Hosting: 1.862

DIPF
Bildungsforschung
und Bildungsinformation

Weitere Schritte

- Erfahrungsberichte der Teilnehmer der FIS Bildung Tagung
- Erfahrungen der DNB im Projekt PETRUS
- Klärung des Budgets
- Auswahl eines Testsystems
- Alternativ: Kooperation mit der TU Darmstadt
- Test eines Systems
- Auswertung und Interpretation der Testergebnisse
- Ggf. Implementierung eines Systems

DIPF
Bildungsforschung
und Bildungsinformation

FAZIT

- Große Datenmengen nicht mehr allein intellektuell zu erschließen
- Automatische Indexierung liefert besseren Recall bei schlechterer Precision
- Grundlage für die Qualität der Indexierung ist die Qualität der Dokumentensammlung zum Zeitpunkt des Trainings
- Problematik der Rechte an Datenmaterial zur Verarbeitung

DIPF
Bildungsforschung
und Bildungsinformation

Danke für die Aufmerksamkeit !

Stellvertretender Leiter IZB
Botte Alexander
Tel: (0)69 24708-330
Mail: Botte@dipf.de

FIS Bildung
Wicker Katrin
Tel: (0)69 24708-329
Mail: Wicker@dipf.de

Koordination
Wissel Verena
Tel: (0)69 24708-301
Mail: Wissel@dipf.de

Pedocs
Paulokat Ute
Tel: (0)69 24708-318
Mail: Paulokat@dipf.de

DBS
Massar Tamara
Tel: (0)69 24708-322
Massar@dipf.de

IuD
Balazs-Bartesch, Gerda
Tel: (0)69 24708-426
Mail: Balazs@dipf.de

IuD
Kiersch Almut
Tel: (0)69 24708-308
Mail: Kiersch@dipf.de

Christa Schöning-Walter

PETRUS

Prozessunterstützende Software für die digitale Deutsche Nationalbibliothek

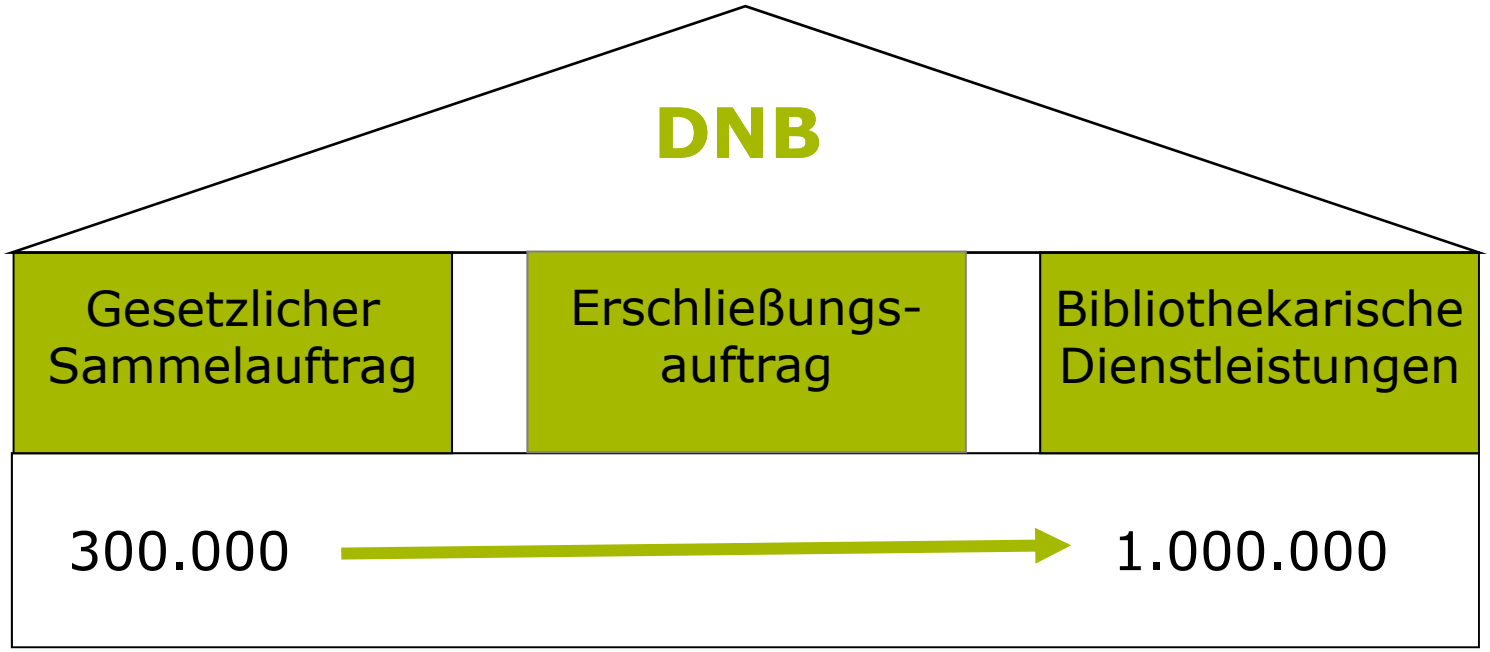
Projektziel

Erarbeitung, Erprobung und Implementierung eines Stufenmodells zur weitgehend automatischen Erschließung für letztlich alle Medienwerke im Sammelauftrag der DNB

Einsatz softwaregestützter Verfahren für

- die automatische Indexierung und Klassifizierung
- die Metadatenextraktion und Metadatengenerierung

Erschließungsvolumen pro Jahr



Leitlinien für PETRUS

- Verknüpfung konventioneller und maschineller Verfahren bei der Formalerschließung und Inhaltserschließung
- automatisierte Verfahren als Basisform der Verarbeitung für alle maschinenlesbaren Objekte
- Option auch für die traditionellen Medien
- die Regelwerke und bibliografischen Dienstleistungen bleiben im Grundsatz unangetastet
- perspektivisch soll auch die Beitragsebene mit in die Erschließung einbezogen werden: Zeitschriften-, Zeitungs-, Sammelbandartikel etc.
- das Finden von Informationen rückt in den Vordergrund: Schaffung neuer Zugänge/Sucheinstiege für den Nutzer

Ergebniserwartung

- modular zusammengestelltes, nachregelbares Gesamtsystem
  - das in Szenarien erprobt und eingeführt wird
  - das mit Differenzierungsstufen für verschiedene Objekttypen neue Geschäftsprozesse und Verfahren unterstützt
- Bewertungssystem mit definierten Qualitätskriterien für die Messung von Qualitätsstufen
  - das als Handlungs- und Entscheidungsbasis und zur Steuerung und Rückkopplung der Prozesse dient

PETRUS-Team

DNB-Standorte

Leipzig

Frankfurt am Main

Berlin

Vorgehen in den PETRUS-Szenarien

- Festlegung der Anforderungen und Qualitätskriterien (fachlich/technisch)
- Aufbau/Auswahl des Testkorpus
- Entwicklung der Teststrategie (Testfälle)
- Aufbau der Testumgebung
- Durchführung von Testläufen
- Auswertung und Diskussion der Ergebnisse
- Festlegung der Anwendungsbereiche
- Entwicklung, Erprobung und Einführung der Geschäftsprozesse

Laufende Szenarien (1/2)

Normdatenrelationierung:

- Übernahme von Fremddaten für die automatische Erstellung von Titeldatensätzen und Normdatensätzen
- Generierung von Verknüpfungen zwischen den Normdaten und den Titeldaten

Paralellausgaben:

- automatische Verknüpfung von Parallelausgaben
- maschinelle Übernahme bereits vorhandener Erschließungsdaten in die Parallelausgabe

Laufende Szenarien (2/2)

automatische Vergabe der DNB-Sachgruppen:

- Einordnung in die 100 Sachgruppen (Kategorisierung)

automatische Beschlagwortung:

- Sachschlagwörter aus der SWD als kontrolliertes Vokabular
- zusätzliche Vergabe freier Deskriptoren
  (als weitere Sucheinstiege und zur Unterstützung der SWD-Pflege)
- zukünftig evtl. weitere Schlagwörter (z. B. Geografika), andere Thesauri (z. B. STW) etc.

Derzeitiges Testkorpus

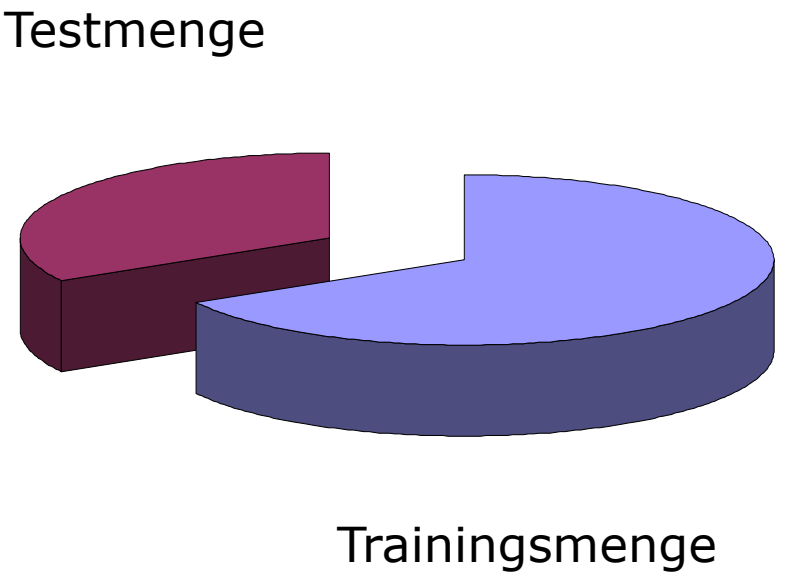
gescannte Inhaltsverzeichnisse (deutsch, DNB + HEBIS):

Umfang: ca. 120.000 Objekte

Volltexte (deutsch):

Online-Hochschulschriften
+ Online-Monografien

Umfang: ca. 45.000 Objekte



Ausschreibungsverfahren

„Erprobung softwaregestützter Verfahren für eine automatische Verschlagwortung und Klassifizierung“

- paralleler Test verschiedener Systemlösungen
- Veröffentlichung der europaweiten Ausschreibung am 24.09.2009
- Forderung: Überlassung geeigneter Softwaresysteme für zeitlich befristete Funktionstests zur
  - automatischen Beschlagwortung (unter Einbindung eines vorgegebenen kontrollierten Vokabulars)
  - automatischen Klassifizierung (Kategorisierung nach einem vorgegebenen Schema)

Erfolgreiche Anbieter

- Intrafind, Planegg bei München
  - Intrafind TopicFinder
- iSquare, Berlin
  - iSquare SmartSearch
- Averbis, Freiburg
  - Averbis Extraction Platform
- Rapid-i, Dortmund
  - RapidMiner

Automatische Sachgruppenvergabe

- automatische Zuordnung zu einer Hauptsachgruppe (Option bei Bedarf: Vergabe von bis zu 3 Sachgruppen)
- Steuerung über den Konfidenzwert möglich?

Qualitätsziel:

Die Hauptsachgruppe soll in mindestens 80 % der Fälle richtig zugeordnet werden (Maßstab ist die intellektuell vergebene Sachgruppe).

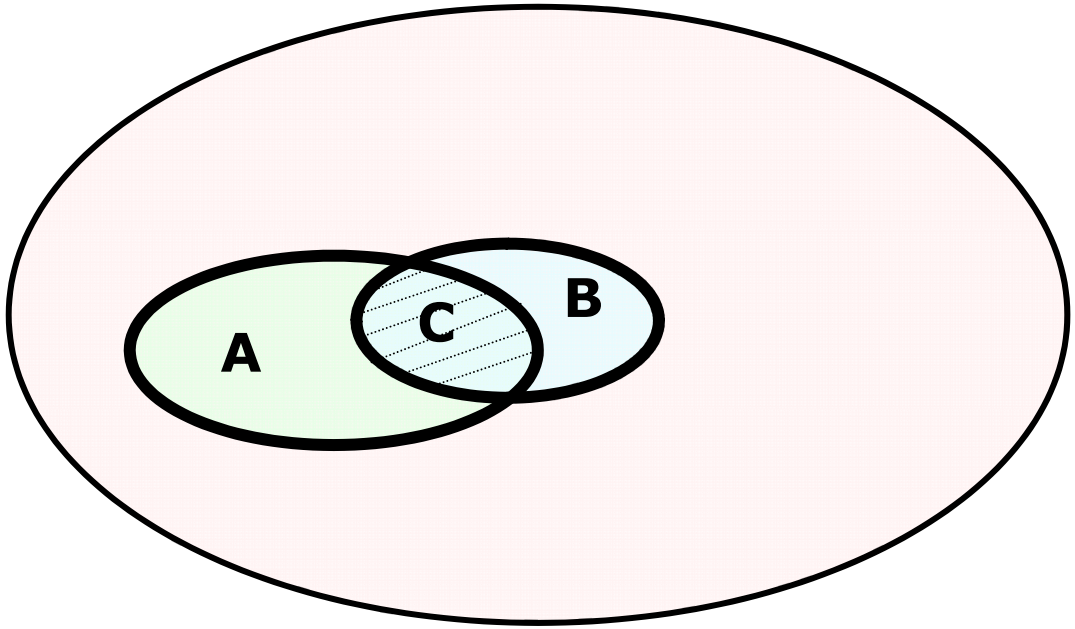
Systematik der DNB-Sachgruppen

- Seit dem Bibliografie-Jahrgang 2004 werden die Reihen der Deutschen Nationalbibliografie nach 100 Sachgruppen gegliedert:
  - System zur thematischen Ordnung von Titeldaten unabhängig von der Sprache der Publikationen
  - beruht auf der Dewey-Dezimalklassifikation (DDC)
  - weitere Anwender: Österreichische Bibliografie, Schweizer Buch (mit jeweils nur geringfügigen Unterschieden)
  - Anwendungszwecke: Gliederung von Nationalbibliografie und Neuerscheinungsdienst, Browsingfunktionen im Portal, Geschäftsgangsteuerung etc.

Maße zur Beurteilung der Erschließungsergebnisse

$$\text{recall} = \frac{C}{A}$$

$$\text{precision} = \frac{C}{B}$$



Kombination der beiden Maße:

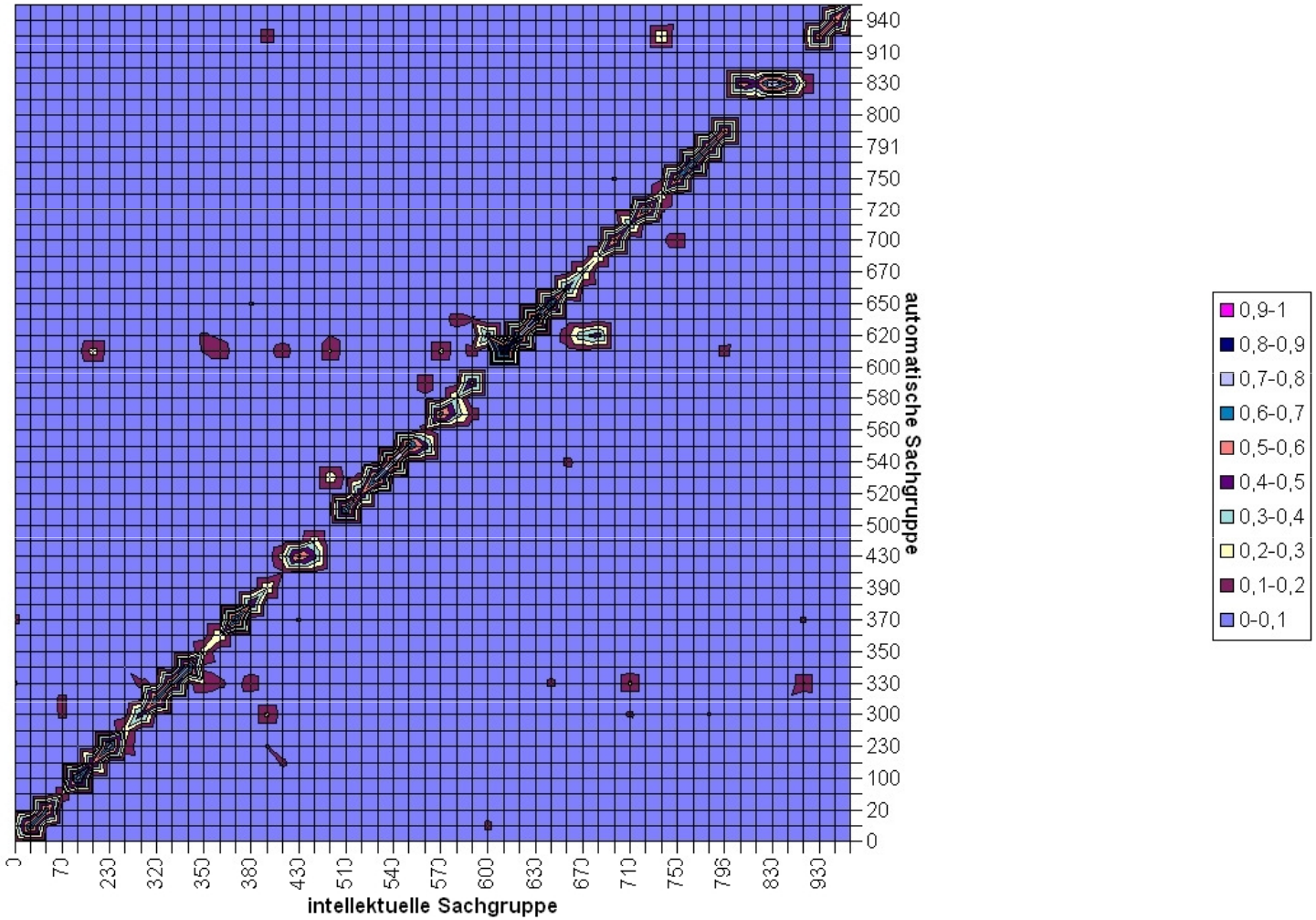
F-measure (Harmonisches Mittel)

$$F = \frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

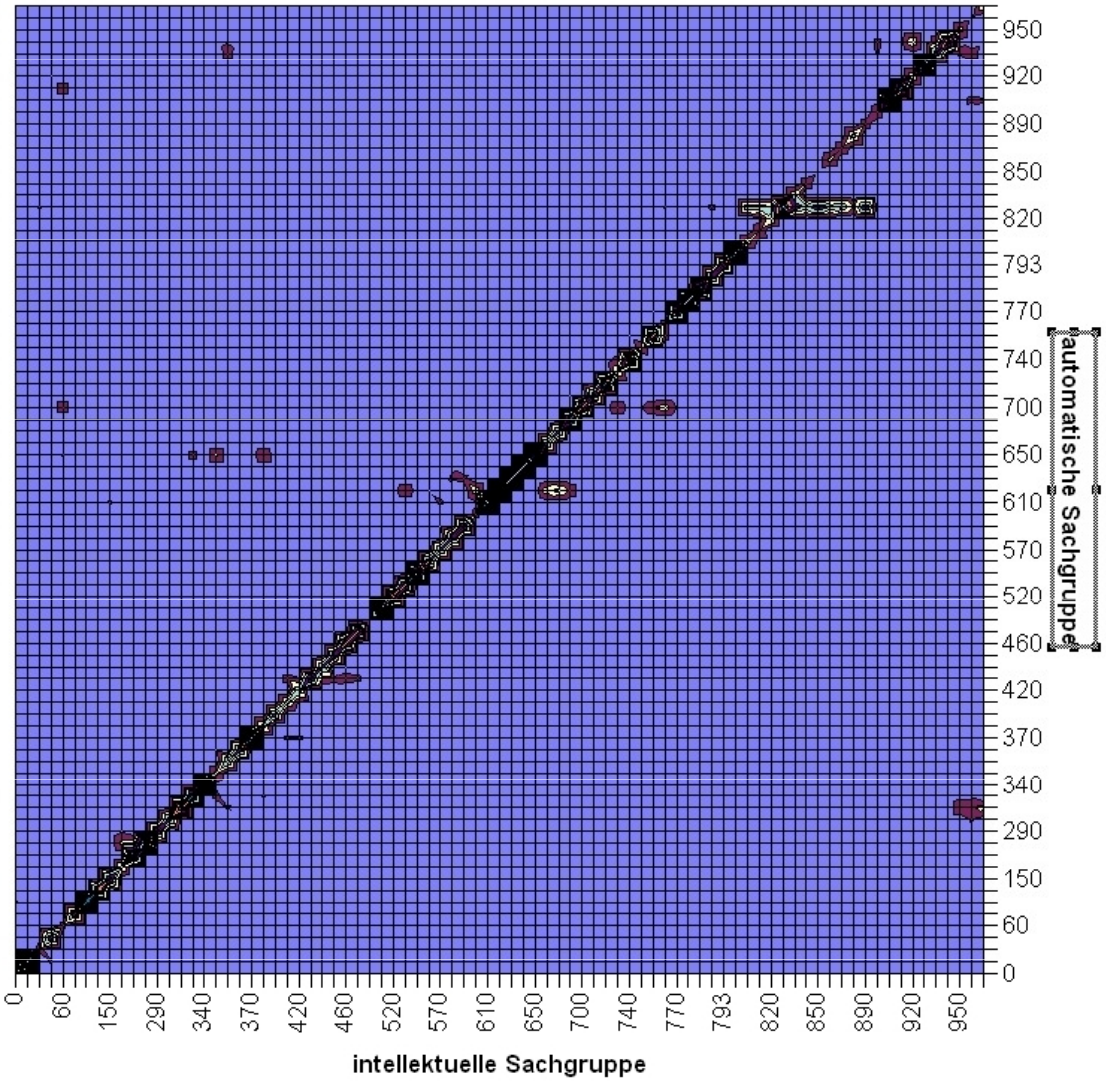
Herausforderungen und Schwierigkeiten

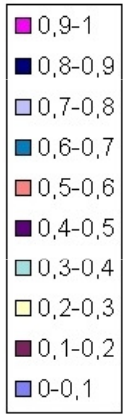
- Sachgruppen, die sich nicht deutlich abgrenzen
- Sachgruppen mit wenigen Dokumenten (nicht ausreichend für das Training)
- ungleichmäßige Verteilung der Objekte auf die Sachgruppen
- neue Sachgruppe ab 2010:
  - 333.7 Natürliche Ressourcen, Energie und Umwelt
- heterogene Objekte: Thema, Textlänge etc.
- Fehler bei der Formatkonvertierung

Testbeispiel Online-Volltexte



Testbeispiel Inhaltsverzeichnisse





Automatische Beschlagwortung

- Ausgabe von SWD-Sachschlagwörtern mit Konfidenzwert
- Ausgabe von freien Deskriptoren mit Konfidenzwert

Anforderungen:

Die Anzahl der Schlagwörter wird nicht vorgegeben, sondern über einen geeigneten Schwellenwert gesteuert.

Die erste Stufe des Thesaurus enthält (nur) die Sachschlagwörter der SWD (160.000).

Beurteilung der Erschließungsqualität über Stichproben

	essentiell (Volltreffer, oder z.B. verwandter Begriff)	nützlich (z.B. direkter OB, direkter UB)	wenig nützlich (z.B. entfernter OB, entfernter UB)	falsches Homonym	anderweitig irreführend
Schlagwort 1					
Schlagwort 2					
...					

Kontakt

Deutsche Nationalbibliothek

Christa Schöning-Walter

Digitale Dienste

+69-1525-1014

c.schoening@d-nb.de

http://www.d-nb.de/

ZBW
Deutsche Zentralbibliothek für Wirtschaftswissenschaften

Workshop - „Unterstützung der Dokumentenerschließung durch automatisierte Verfahren“

Manfred Faden (Hamburg)
Thomas Groß (Kiel)

FFM, 04.05.2010

Was Sie erwartet

1. Standard-Thesaurus Wirtschaft (STW)
2. Motivation
3. Kurzvorstellung der Pilotprojekte
4. MindServer-Komponenten
5. Projektauftrag
6. Stand der Dinge
7. Statistische Auswertung des Pilotprojektes
8. Fazit & nächste Schritte

1. Standard-Thesaurus Wirtschaft (STW)

Polyhierarchischer Thesaurus mit sieben Subthesauri, ca. 6.000 Deskriptoren in dichter semantischer Relation und ca. 24.000 Verweisen als zusätzliches Einstiegsvokabular.

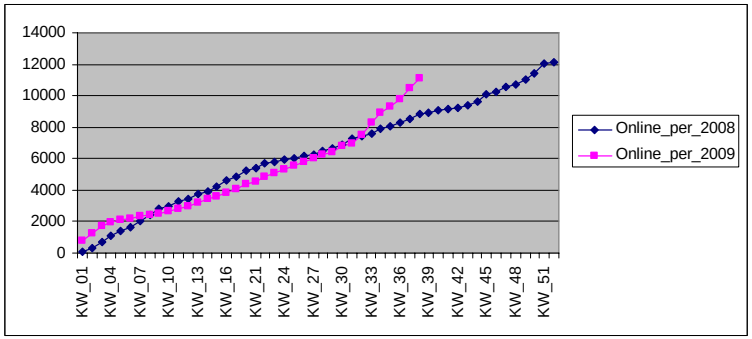
Seit der Version 8.03 ist jeder Deskriptor mit einer englischen Vorzugsbennenung versehen und im Laufe der Zeit wurden und werden weitere englische Verweise mit in das Vokabular aufgenommen.

Seit der Version 8.04 ist der STW unter einer Creative Commons Lizenz online zum Download bereitgestellt: http://zbw.eu/stw

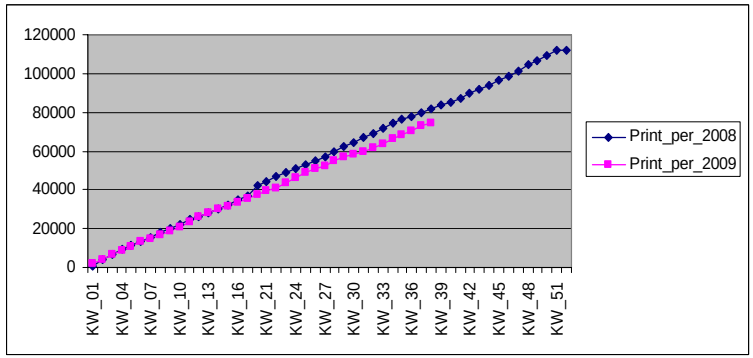
Die aktuelle Version des STW ist 8.06.

2. Motivation: Input 2008-09

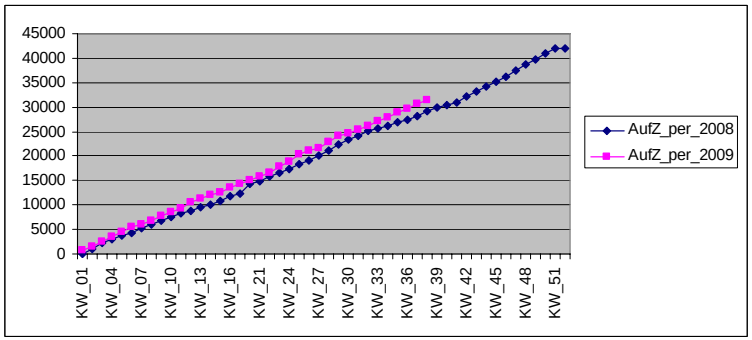
Online



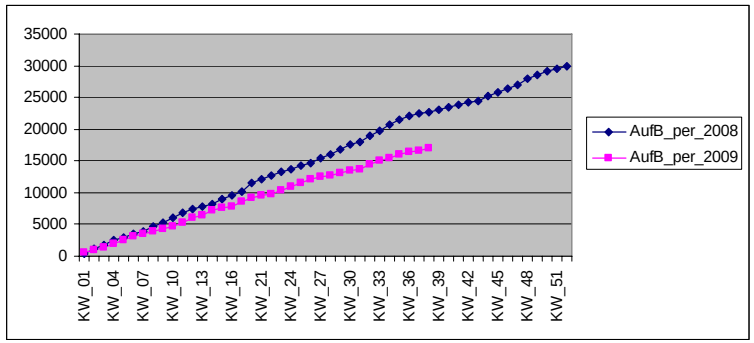
Print



Zeitschriftenaufsatz



Buchaufsatz



3. Pilotprojekt 1 & 2

Pilotprojekt 1:

Ca. 120.000 Datensätze aus Econis und Econstor

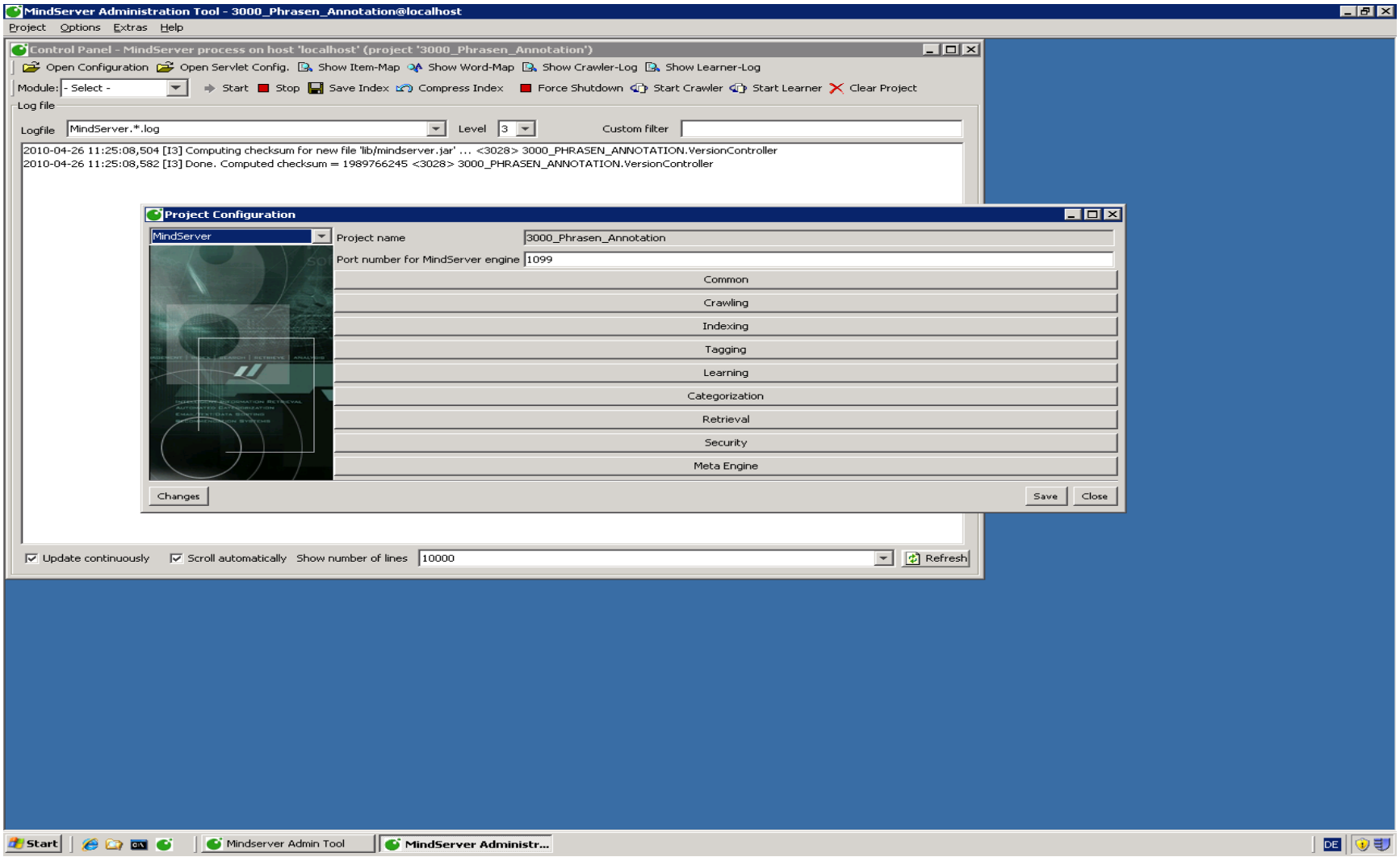
Qualifikation: Alle Aufnahmen, die eine URL aufwiesen.

Pilotprojekt 2: Ca. 40.000 Datensätze

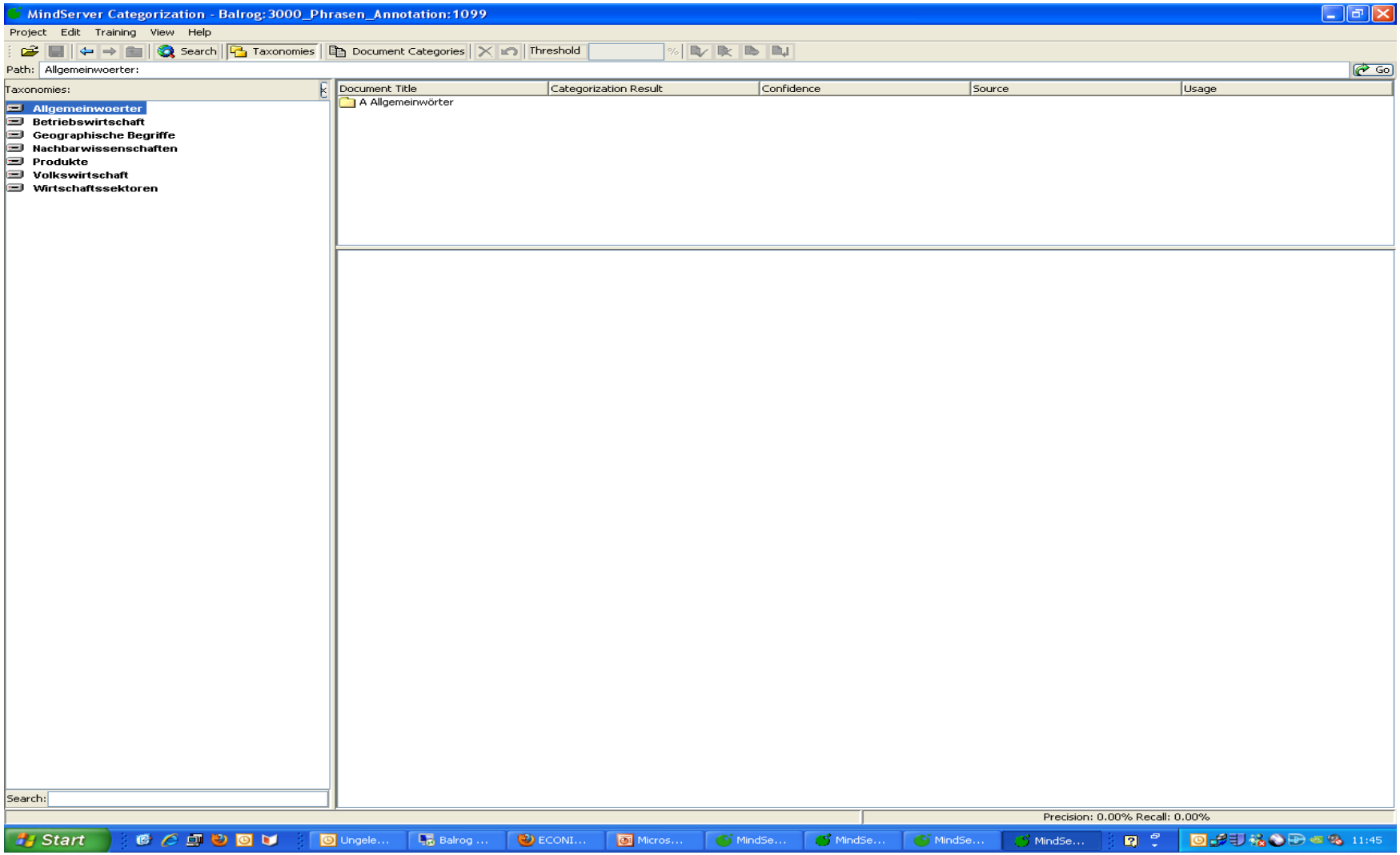
Qualifikation: Alle Datensätze aus 1., deren Text gecrawlt werden konnte

Dazu mehr von Herrn Groß im zweiten Teil der Präsentation (Auswertung, Teil 6)

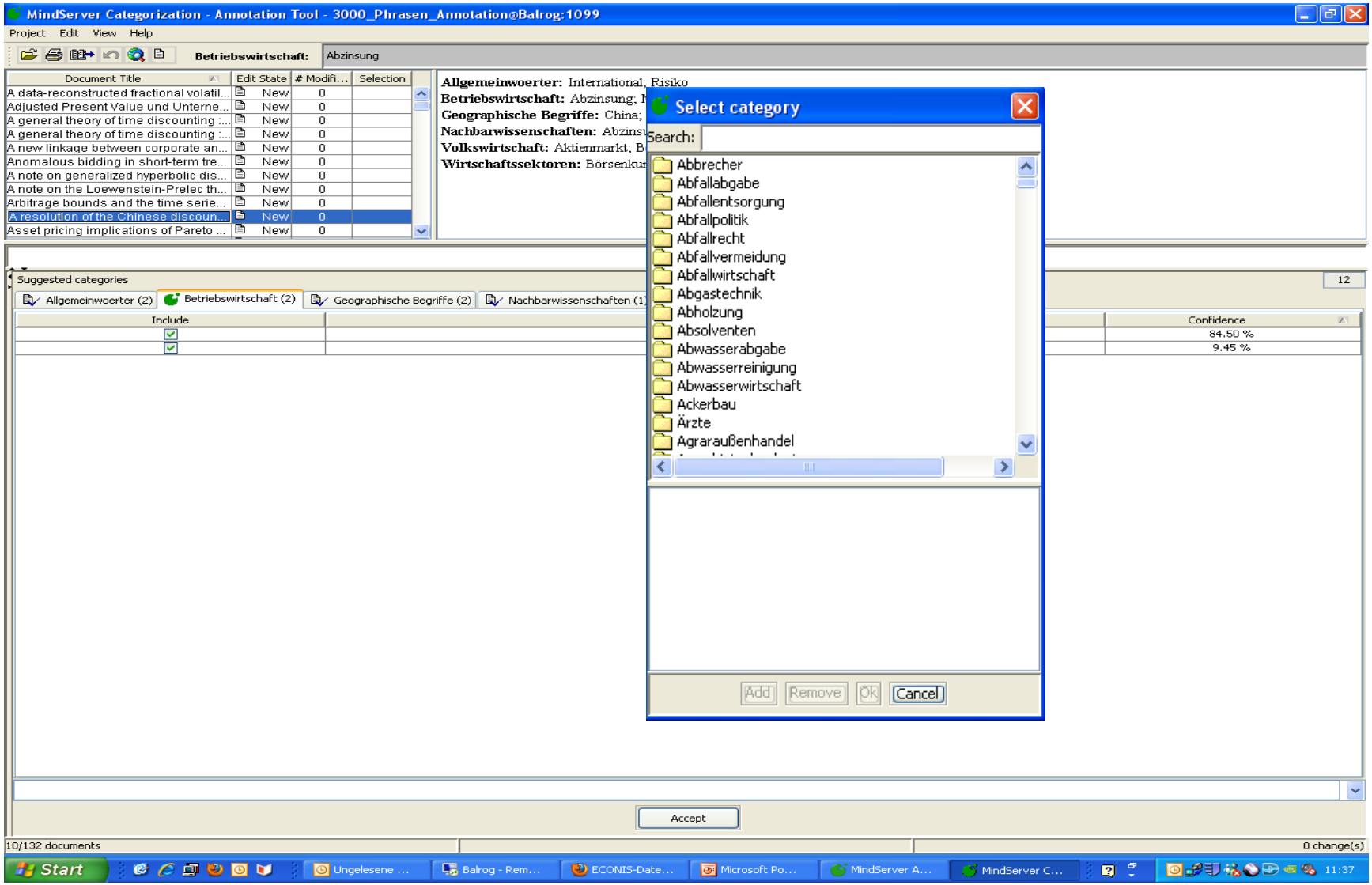
4. MindServer-Komponenten I - Administrationstool



4. MindServer-Komponenten II - Taxonomy Browser

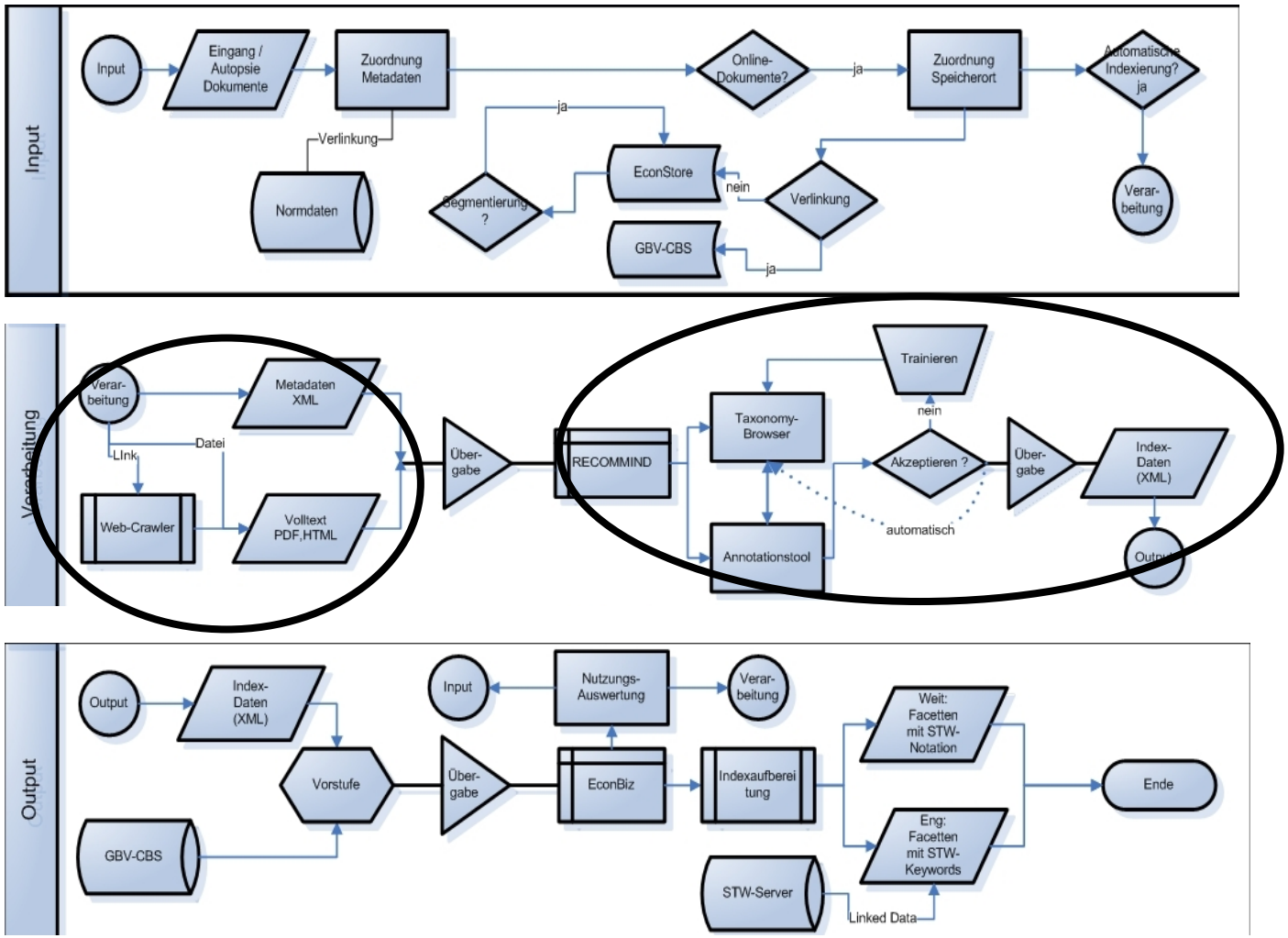


4. MindServer-Komponenten III - Annotationstool



5. Automatische Indexierung – Projektauftrag

Max-Michael Wannags 13.07.2009



6. Stand der Dinge

Stand:

Pilotprojekte: April – November 2009

Entscheidung für einen Kauf des MindServer: Dezember 2009

MindServer im Hause installiert: Ende März 2010

Schulung für IT- und Contentadministratoren: April 2010

7. Auswertung – Grundsätzliches I

Evaluierung des 2. Pilotprojektes

- ca. 40.000 Titel
- Qualitative Bewertung der automatisch generierten Sacherschließungsergebnisse
- Begriffsorientiertes Verfahren
- Vergleich Mensch vs. Automatik
- Trainingsset (26.645 Dokumente) und Testset (12.233 Dokumente), auch Unterstichproben (Datenlage)

7. Auswertung – Grundsätzliches II

Ergebnisbewertung: bekannte Verfahren

- Retrievaltest (Recall/Precision)
- Qualitative Einschätzung

Evaluierung in der ZBW

- Kriterienset nach Stock (2000), vgl. Lancaster (2003)
  - Indexierungskonsistenz
  - Indexierungstiefe (Indexierungsbreite, -spezifität)
  - Indexierungseffektivität
- zusätzlich: Belegungsbilanz, ReferentInnenauswertung

7. Auswertung – Ergebnisse I

Indexierungskonsistenz: Mensch vs. Automatik

Die I. misst den Grad der Übereinstimmung unterschiedlicher Indexierungsergebnisse der gleichen Vorlage.

- Minimum: 0% (718 Titel = 2,8% der Gesamttitelzahl)
- Maximum: 100% (90 Titel = 0,3% der Gesamttitelzahl)
- Mittelwert: 36%
- wenig Titel über 80% (nur 300), größtenteils 20-50%

7. Auswertung – Ergebnisse II

Indexierungstiefe

Dasjenige Indexat der gleichen Vorlage mit gleicher Deskriptorenanzahl ist tiefer erschlossen, welches spezifischere Deskriptoren enthält.

	Automatik	Manuell	Automatik bereinigt*
Minimum	0,00	0,00	0,00
Maximum	1,72	1,33	0,67
Mittel	0,28	0,23	0,14

*nach Indexierungskonsistenz

7. Auswertung – Ergebnisse III

Indexierungseffektivität

Die I. ermittelt, wie oft ein Deskriptor im jeweiligen Datensatz vergeben worden ist.

	Automatik	Manuell
Minimum	2,110	3,629
Maximum	12,439	16,247
Mittel	9,062	13,678

7. Auswertung – Ergebnisse IV

Belegungsbilanz

Die B. fokussiert auf die tatsächliche Nutzung des zur Verfügung stehenden kontrollierten Vokabulars (hier: STW).

Gesamtdeskritorenanzahl	5.770	100 %
Manuell:		
Benutzte Deskriptoren	4.106	71 %
Unbenutzte Deskriptoren	1.664	29 %
Vergabehäufigkeit	32	21 %
Automatik:		
Benutze Deskriptoren	1.002	17 %
Unbenutzte Deskriptoren	4.768	83 %
Vergabehäufigkeit	302	20 %

7. Fazit & nächste Schritte

Fazit:

→ intellektuelle und automatische Sacherschließung zu 1/3 deckungsgleich (ohne Training, ohne Regelvergabe)

→ Automatisches Verfahren wählt häufig Oberbegriffe, Allgemeinwörter (Trainingsverhalten)

→ Intellektuelle Indexierung trennschärfer

→ Automatik benutzt nur Bruchteile des STW (vgl. Zipf`s Gesetz)

Nächste Schritte:

- Test verschiedener Möglichkeiten klassifikatorischer bzw. verbaler Erschließung mit den derzeit eingespielten Taxonomien
- Tests mit einer Änderung der Taxonomie (z.B. alphabetisch flach)

Literatur

Stock, Wolfgang (2000): Informationswirtschaft – Management-wissen für Studium und Praxis. München: Oldenbourg-Verlag.

Lancaster, Frederick Wilfried (2003): Indexing and abstracting in theory and pratice. London: Facet.

Vielen Dank für Ihre Aufmerksamkeit

Anhang

Berechnungen

Indexierungstiefe = $\{ld[HE(B_1)+1]+\ldots+ld[HE(B_n)+1]\}/S$

HE = Hierarchieebene des Deskriptors (B), S = Seitenzahl

Indexierungseffektivität = IDF(B) = [ld(N/n)]+1

IDF = Inverse Dokumenthäufigkeit, N = Gesamttitelzahl
n = Titel mit jeweiligem Deskriptor

Indexierungskonsistenz (V1, V2) = A / B+C-A

V1, V2 = Vorlagen (Dokumente), A = übereinstimmende Deskriptoren, B = Indexierungsbreite Vorlage V1, C = Indexierungsbreite V2

gesis
Leibniz-Institut für Sozialwissenschaften

Datenbank SOLIS: Automatische Indexierung mit MindServer - Categorization

Trends aus ersten Analysen

Monika Zimmer

04. Mai 2010

MindServer - Categorization

- ***Trainingsbasis:***
  368.000 Dokumente aus SOLIS
- ***Arbeitsweise / Prinzip:***
  „Probabilistische latente semantische Analyse“ (PLSA) in Verbindung mit „Support Vector Machines“

Intellektuelle vs. automatische Indexierung: ***I.Vergleich Sachgebietsklassifikationen*** (N= 40 Dok.)

- Anzahl Klass. intellektuelle Indexierung i.D. => 2,2
- Anzahl Klass. automatische Indexierung i.D. => 3,9
- Anzahl identischer Klassifikationen i.D. => 1,1
- Hauptklassifikation bei automatischem Verfahren durchschnittlich in 70% der Fälle enthalten

Intellektuelle vs. automatische Indexierung:

II. Vergleich Schlagwörter (N= 40 Dok.)

- Anzahl SW intellektuelle Indexierung i.D. => 17,3
- Anzahl SW automatische Indexierung i.D. => 19,1
- Anzahl identischer Schlagwörter i.D. => 5,8
- Anzahl zusätzlicher SW „passend“ – „verwandt“ bei automatischer Indexierung i.D. => 3,3

Psychologie Information
ZPID – Leibniz Institut

FIS-Bildung Fachtagung 3./4. Mai 2010

Dipl.-Psych. Michael Gerards

Der Einsatz eines automatischen Indexierungssystems im ZPID

Vorschau

- Arbeitsweise von AUTINDEX
- Anpassung von AUTINDEX
- Integration von AUTINDEX
- Evaluation von AUTINDEX

Zielsetzung der automatische Indexierung

Ziel einer automatischen Indexierung:
Unterstützung der intellektuellen Verschlagwortung

dazu seit 2006 eingesetztes Verfahren: AUTINDEX (**Auto**matic **Index**ing) vom "Institut der Gesellschaft zur Förderung der Angewandten Informationsforschung e.V." (IAI)

Arbeitsweise von AUTINDEX 1

Aufgabe: Generierung von Deskriptorvorschlägen aus Dokumenten (Titel, Abstract, Autorenschlagworte)

Vorgehen von AUTINDEX: Einsatz linguistischer Intelligenz bei der Textanalyse kombiniert mit statistischen Elementen

- AUTINDEX arbeitet "verstehensbasiert" und nicht "string-basiert"
- "Ausreizung" der linguistischen Möglichkeiten, um statistische Verfahren für die Indexierung optimal zu unterstützen

Arbeitsweise von AUTINDEX 2

Linguistische Analyse mit u.a.

- Satz- und Wortformerkennung
- Erkennung von morphologischen (Haus/Häuser), syntaktischen (Reduktion von Kosten -> Kostenreduktion) und semantischen (Synonyme) Varianten von Worten
- Kompositazerlegung
- Ausschluss von Stopwörtern und ihren morphologischen Varianten
- Zuordnung des ermittelten Begriffs zu einer semantischen Klasse

Arbeitsweise von AUTINDEX 3

Nach der Bearbeitung linguistischer Einheiten: Gewichtung der ermittelten Terme unter Berücksichtigung der

- Häufigkeit des Vorkommens des Terms im Text (Termfrequenz)
- Stellung des Terms im Text
- im Text vorkommenden semantischen Klassen

Ergebnis: Wörter und Phrasen, die häufig bzw. an prominenter Stelle im Text vorkommen bzw. zu den häufigsten semantischen Klassen gehören, werden als Deskriptorkandidaten gesammelt und mit den Begriffen aus dem Thesaurus abgeglichen.
Wenn Übereinstimmung gefunden und ein definierter Schwellenwert überschritten wird, erfolgt Deskriptorvorschlag.

Arbeitsweise von AUTINDEX 3

Voraussetzung für die linguistische Analyse:

Fachsprache muss in das AUTINDEX-Wörterbuch integriert sein

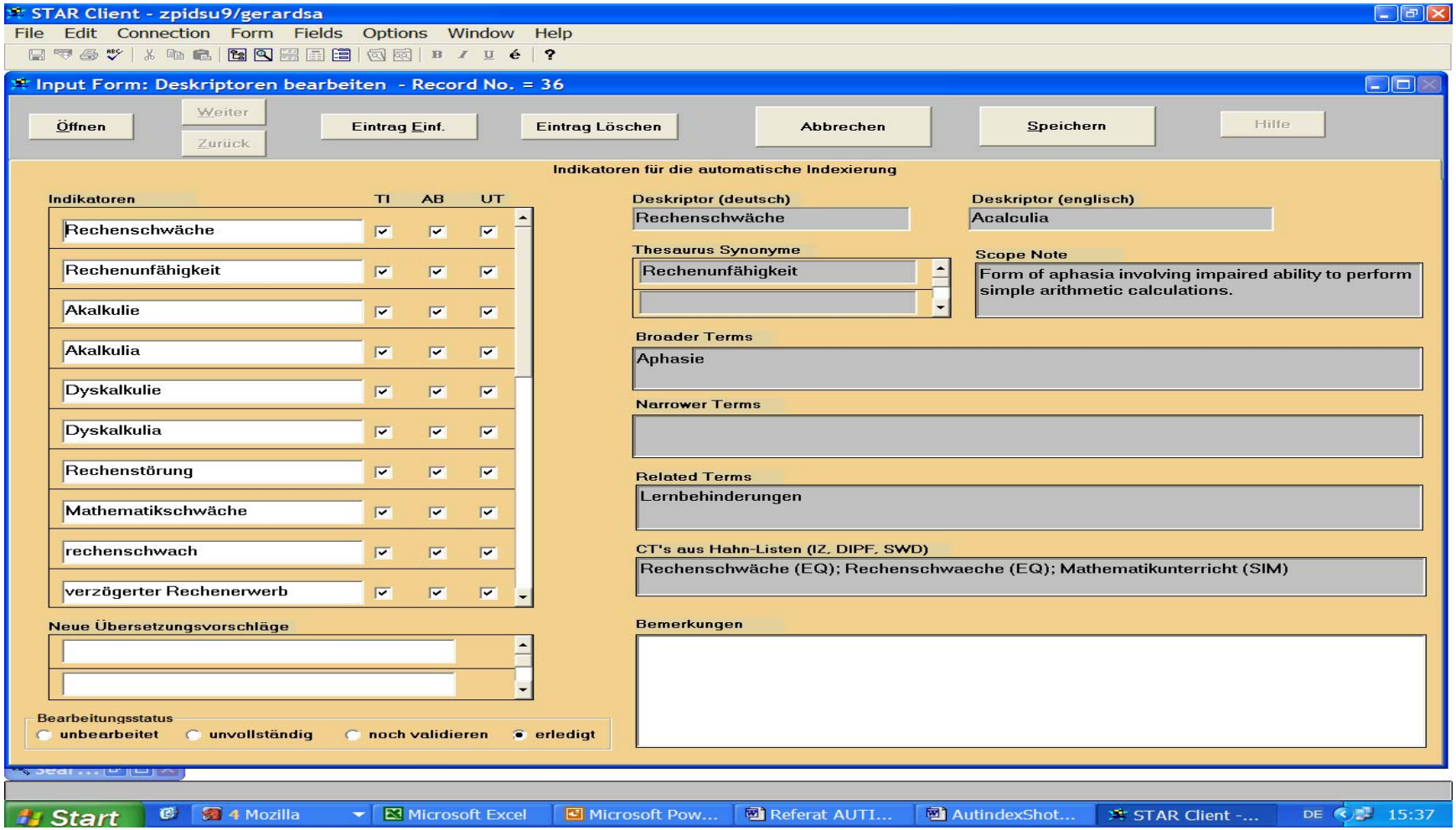
- Einbau der Deskriptorterme und der im Thesaurus enthaltenen Synonyme (ca. 6700 Begriffe)

da unzureichend

- Entwicklung des Indikatorkonzeptes: Einbau von Begriffen, die in enger Beziehung zu den Deskriptoren stehen, ohne als direkte Synonyme zu gelten (ca. 24.000 Begriffe)

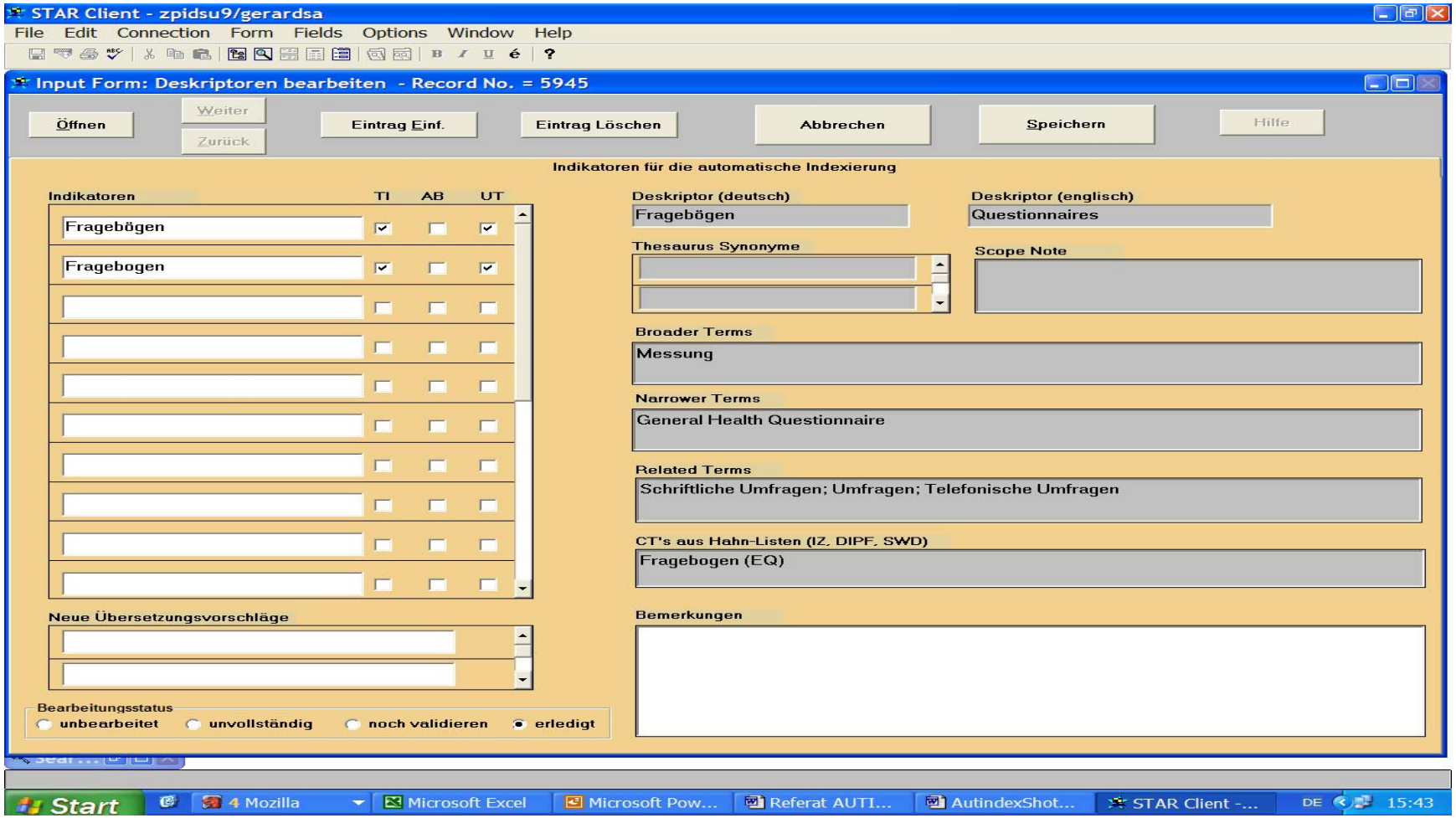
Erweiterung des Thesaurus durch Indikatoren

Beispiel Rechenschwäche (Acalculia)



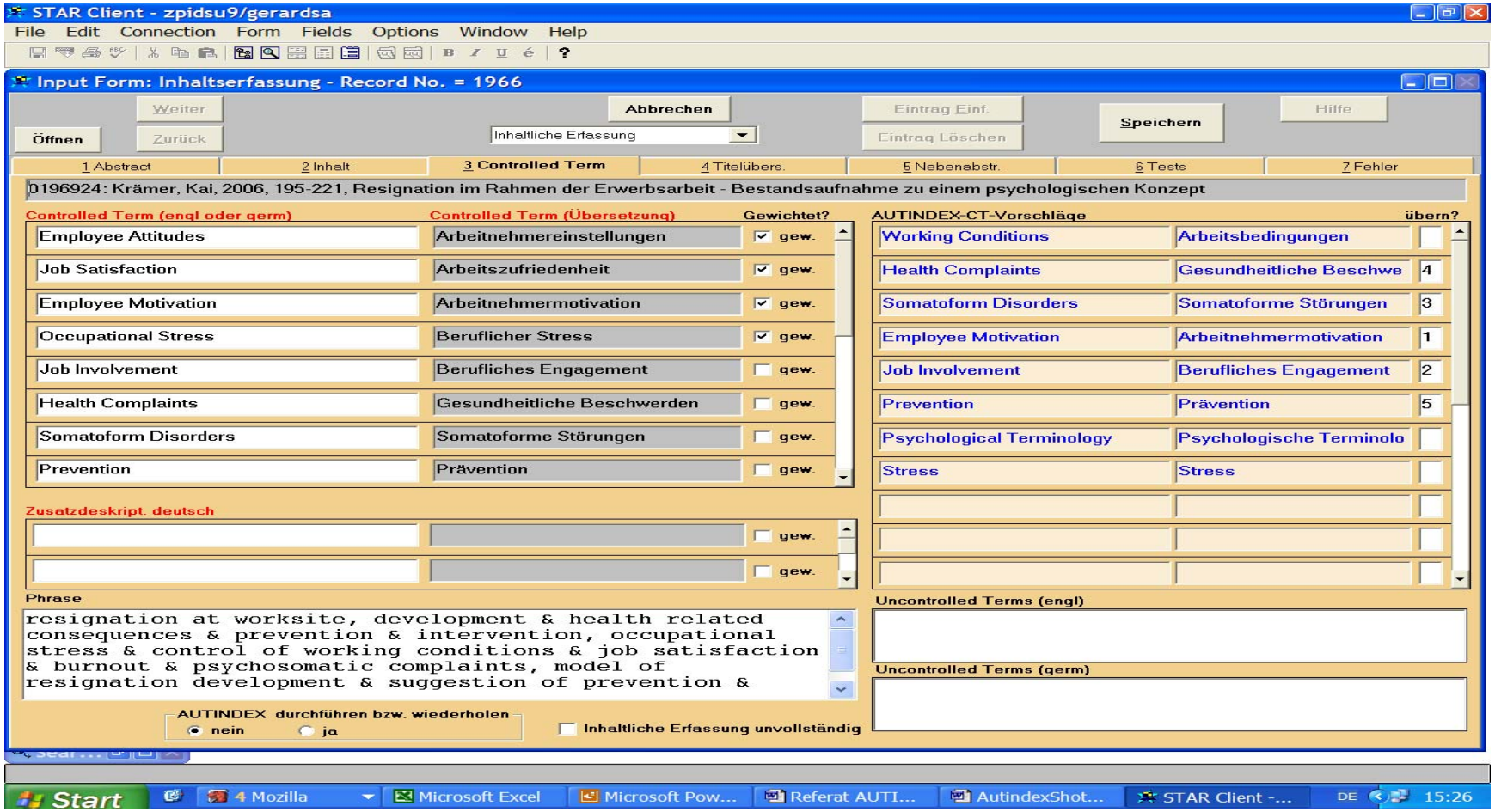
Eliminierung ungeeigneter Thesaurusbegriffe

Beispiel Fragebögen (Questionnaires)



Integration von AUTINDEX

Übernahme von Deskriptorvorschlägen



Evaluationsstudien - Vorgehensweise

- Vergleich von AUTINDEX versus Intellektuelle Indexierung in 2 älteren Studien mit jeweils 63 Dokumenten aus PSYNDEX
- Studie 2010: Wie viele Deskriptoren werden von AUTINDEX übernommen?
- Konsistenzindex zum Vergleich der verschiedenen Studien
- Vergleich DIPF / IZ / ZPID

Evaluationsstudie 2010 - Übereinstimmungen

3016 Dokumente (Zeitschriftenaufsätze)	Anzahl der Deskriptoren	Durchschnitt
Intellektuell	24532	7,8
Automatisch	21789	7
Übereinstimmungen	8782	2,8
Für 3 % der Dokumente werden keine Vorschläge gemacht		

Evaluationsstudie - Konsistenzindex nach Rolling

Studie	Konsistenz
Automatisch-Intellektuell 2004 (63 Dok., ohne Indikatoren)	25 %
Automatisch-Intellektuell 2006 (63 Dok., mit Indikatoren)	40 %
Automatisch-Intellektuell 2010 (3016 Dok., mit Indikatoren)	38 %
Intellektuell-Intellektuell (37 Dok., Interindexer-Konsistenz)	57 %

Evaluationsstudie 2010 – DIPF / IZ / ZPID

Zeitschriften-aufsätze (06-10)	DIPF 125 DE	IZ 360 DE	ZPID 2531 DE
Intellektuell	6,5	6,9	8
Automatisch	6	6,1	7,2
Überein-stimmungen	2,3	2,5	2,9
Rolling Konsistenz-Index	37 %	38 %	38 %

Vielen Dank für Ihre Aufmerksamkeit.

Michael Gerards

AUTINDEX-Textanalyse