Westfälische
Wilhelms-Universität
Münster

Workshop c3m-II

Information Retrieval

Martin Juhrisch

Projekt MIRO, Universität Münster

Arbeitspaket Information Retrieval
Schild, Perske, El Wardi, Przibytzin

IKM service
ULB • ZIV • UniV

*„Informationen **treffsicher**, **verlässlich**, **vollständig**, **aktuell** und **leicht zugänglich** am Arbeitsplatz des Nutzers verfügbar zu machen“*

- **Beschreibung und Diskursbereich**
- **Aspekte der Evaluierung**
- **Diskursbereich für die Implementierung**
- **Fazit zum Stand des AP7**

Beschreibung

Ausgangslage an der Universität Münster

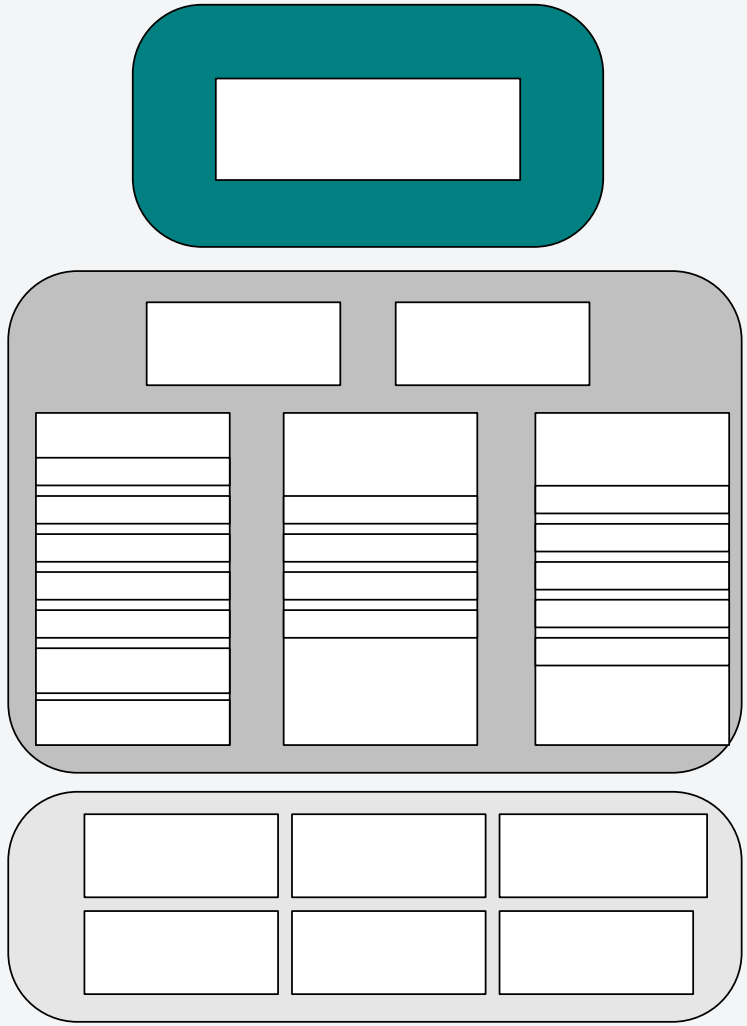
- beschränkter Zugang zu organisatorischen und wissenschaftlichen Informationen
- Schlechte Qualität der Suche in relationalen Datenbanken und Metasuchsystemen

Motivation für den Einsatz dieser Technologie an der WWU

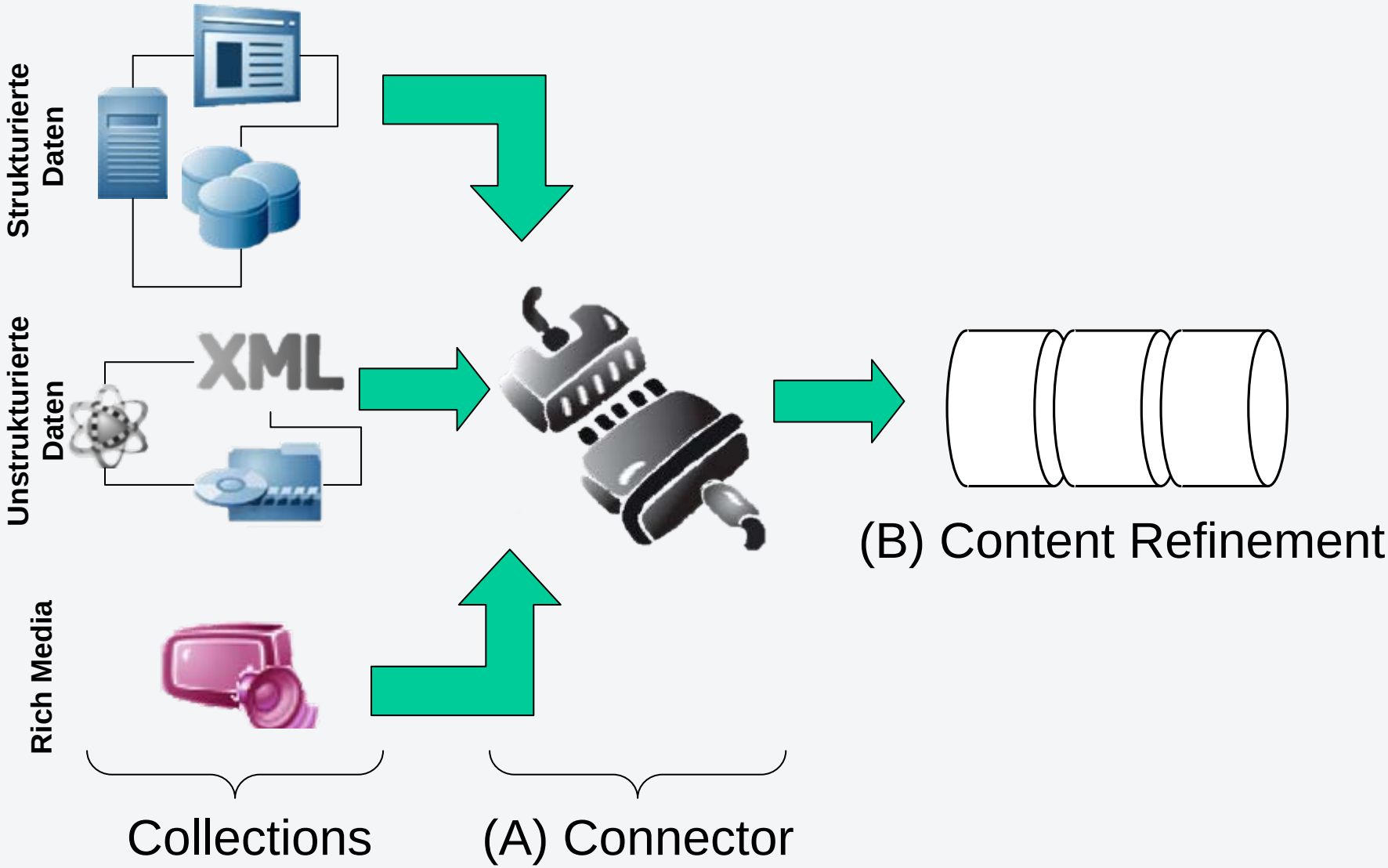
- Vereinheitlichung der wissenschaftlichen Informationsversorgung (Aufgreifen der Ansätze aus Vascoda, BASE)
- Vereinheitlichung des Zugriffs auf heterogene Ressourcen
  - Verbesserung der wissenschaftlichen Informationsversorgung
  - Verbesserung der organisationalen Informationsversorgung
- Verbesserung des Suchkomforts im Vergleich zum Ist-Zustand
  - Rechtschreibvorschläge, Ähnlichkeitssuche, linguistische Verfahren, etc.

Aspekte der Evaluierung

- **Integration in ein Portal**
- **Zugriffssicherung**
- **Sophisticated Features**
- **Security (Identity Management TIM)**
  - **Integration in das Identitätsmanagement**
  - **Single-Sign-On**
  - **Realisation innerhalb der SE oder aufgesetzt**
- **Datenquellen**
  - **Frei konfigurierbare Anbindung von Datenquellen**



Ablauf der Anbindung der Datenquellen



Information Retrieval
(A) **Aspekt der Anbindung von Datenquellen**

CONNECTORS

Notes Exchange DB2 Documentum FileNet PCDocs ODBC Oracle SAP Siebel SharePoint Sybase Moreover POP3 SQL NewsEdge NNTP FTP HTTP FileSystem OpenText Vignette + 200 others

XML

- Datenbankanbindungen
- Erkennung von allen Texttypen
- Vorgefertige Lösungen
- „Leichte“ Konfigurierbarkeit
- Möglichkeit zur Federated Search

Datenbanken

Portale/Web

Legacy

Applikationen

CMS, DMS, Filesysteme

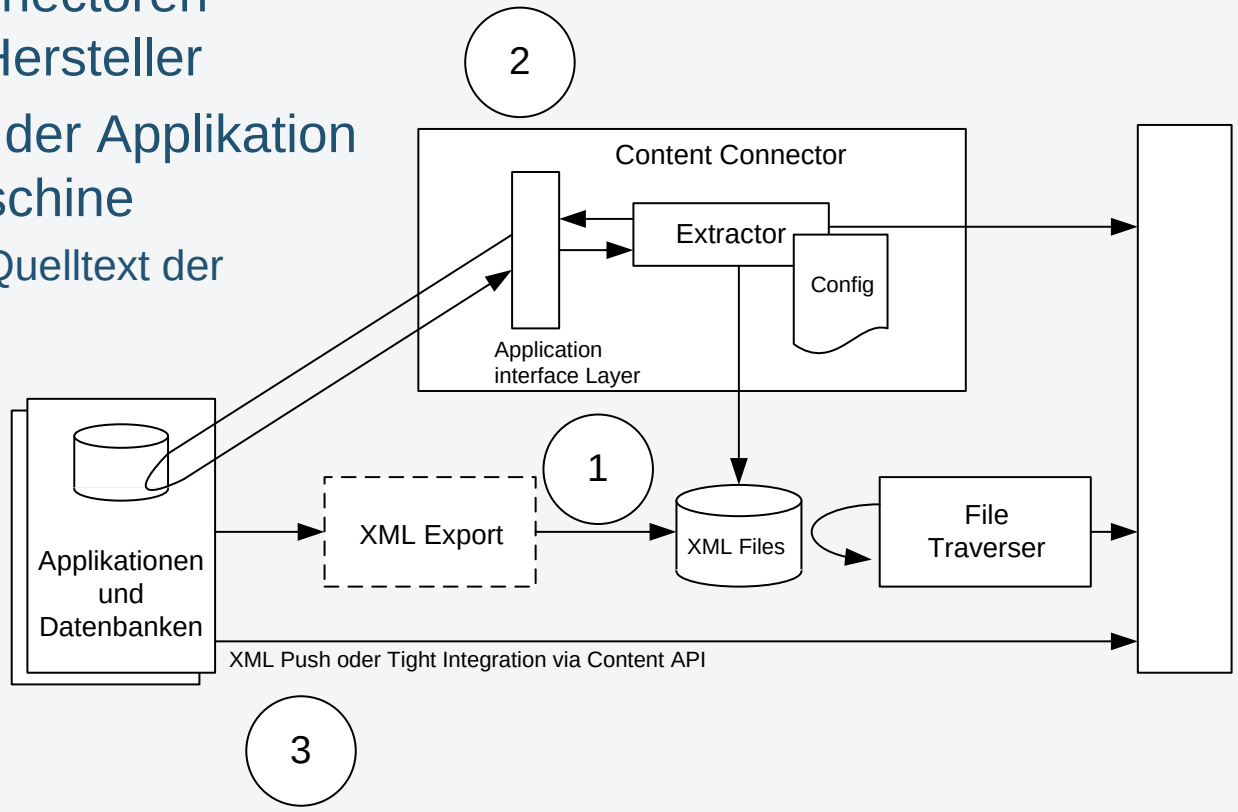
Video/Audio

News

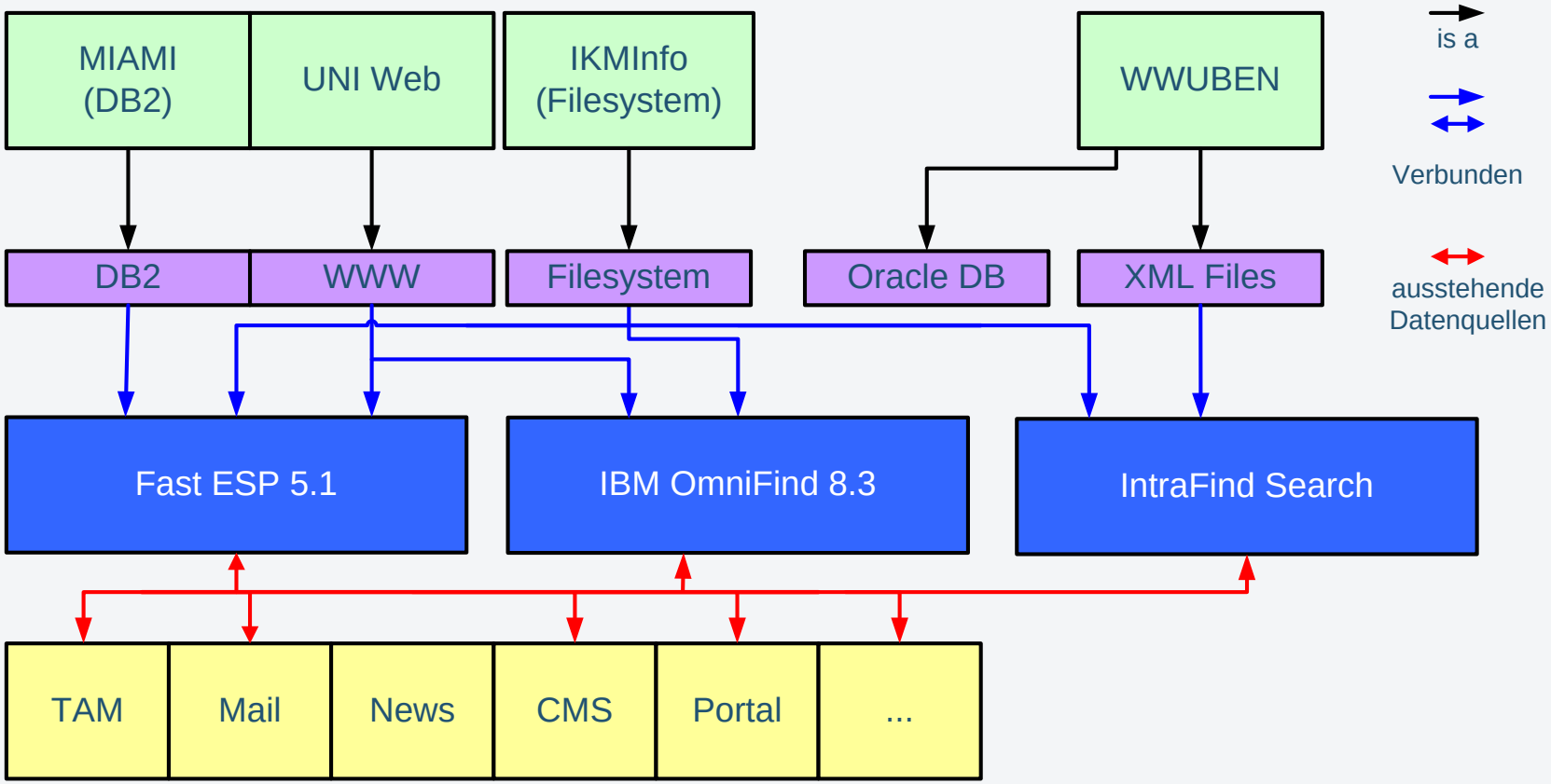
Mail

Information Retrieval
(A) **Möglichkeiten der Anbindung von Datenquellen**

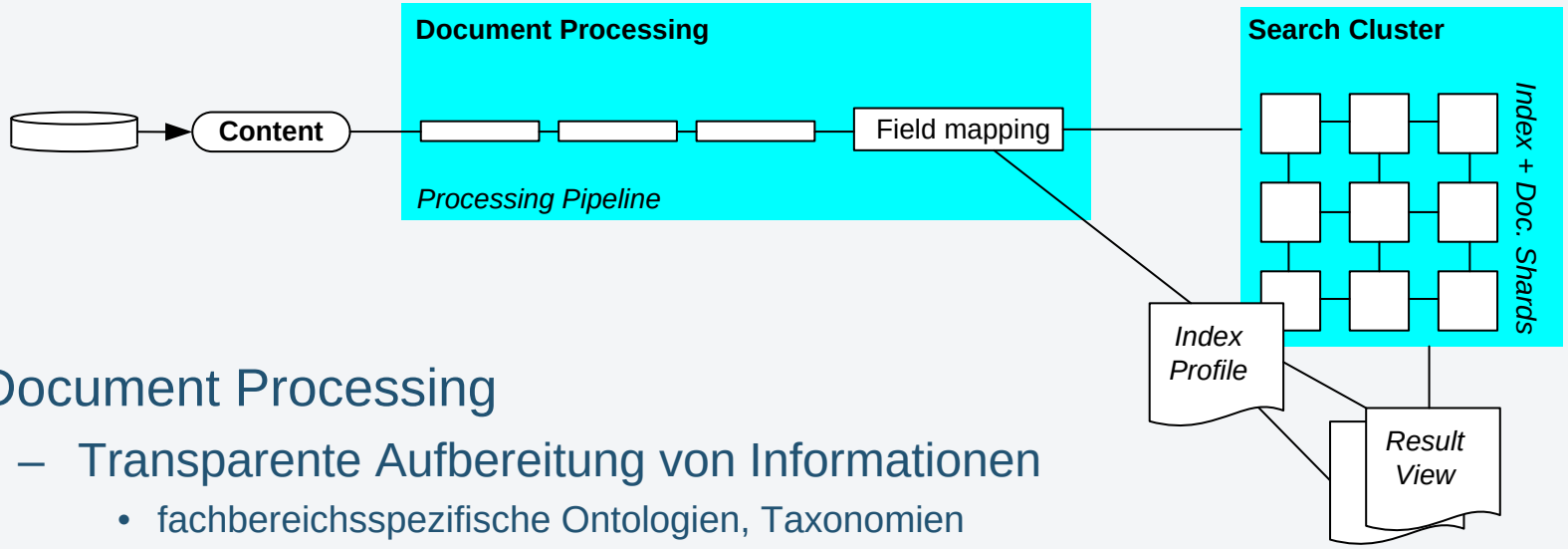
1. Datenquelle exportiert das gesamte Datenvolumen als XML Files
2. Proprietäre Connectoren verschiedener Hersteller
3. Enge Kopplung der Applikation an die Suchmaschine
   – Eingriff in den Quelltext der Anwendung



Information Retrieval
Angebundene Datenquellen



Information Retrieval
(B) **Content Refinement**



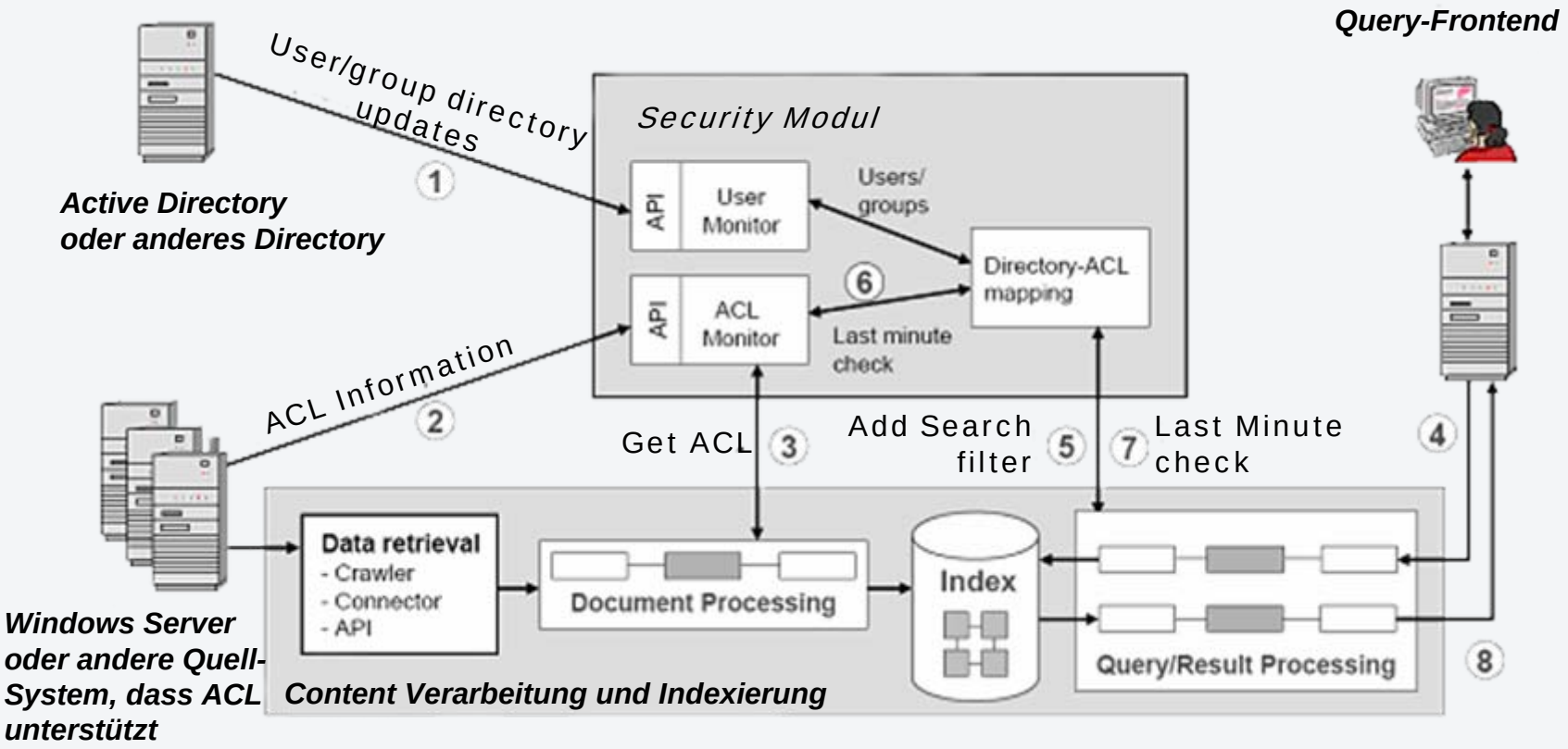
- Document Processing
  - Transparente Aufbereitung von Informationen
    - fachbereichsspezifische Ontologien, Taxonomien
    - Eigene Lexikaentwicklung

- Query Processing
  - Spellchecking – Phonetic match
  - Topic classification
  - Ermittlung von Ranking-Informationen
  - Phrasing etc.

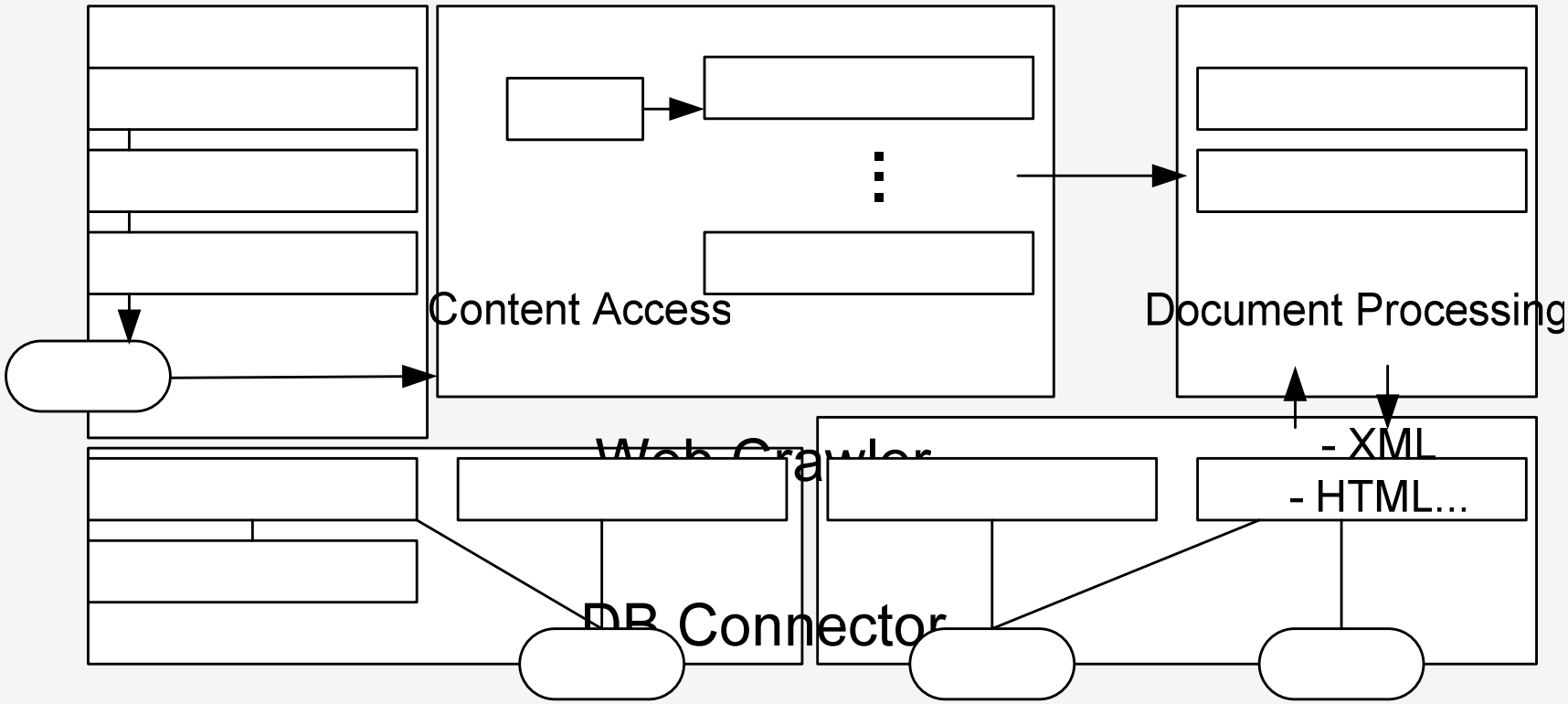
Information Retrieval
Aspekte der „Sophisticated Features“

- Kategorien
  - Vorgefertigte Kategorien je nach Dokumentengruppe
  - „Anlernbare“ Kategorien
- Suchen nach „ähnlichen“ Worten – ling. Methoden
  - Lemmatisierung
  - Stemming
- Auffinden von ähnlichen Dokumenten
  - Clustering
  - Semantische Netze
- Erzeugen von Zusammenfassungen/Abstracts
- Video/Audio-Indexierung
- Benachrichtigungsdienst
  - Push statt Pull

Information Retrieval
Aspekt der Zugriffssicherheit



Diskursbereich für die Konzeption der WWU Lösung



Fazit:

- Architektur der evaluierten Suchmaschinen ähnelt sich stark
- Einarbeitung in Möglichkeiten der Integration läuft bereits

Fazit

- Unsere Grundbedürfnisse werden von allen Anbietern gedeckt
- Unterschiede in den Suchmaschinen gibt es vor allem im Preis und in den „Sophisticated Features“.
- Komponenten die eine Suche aufwerten sind eher in der Entwicklung. Sie sind nicht enthalten oder müssen teuer zugekauft werden.
- AP7 ist auf einem zügigen Weg, die derzeitige Testphase wird mitgenutzt, um sich in die Produkte und die Suchmaschinentechnologie einzuarbeiten.
- Entscheidung für eine Suchmaschine sollte vorzugsweise im Juli getroffen werden.

Vielen Dank

für

Ihre Aufmerksamkeit!