

# Entwicklung eines Dokumenten- Management-Systems

Kolloquium

Vortrag: Jan Löffler

17.12.2003

# Gliederung

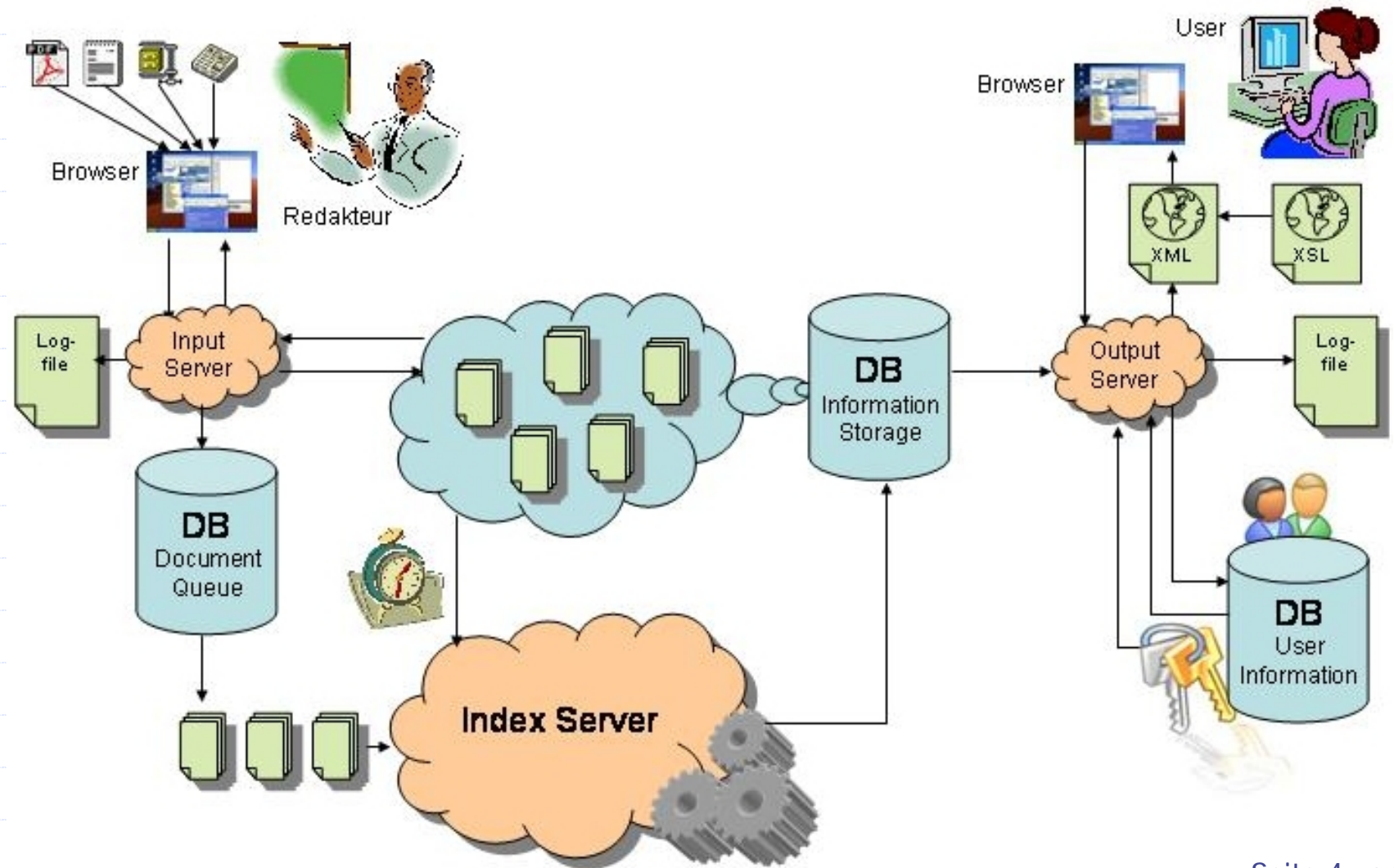
---

1. Beschreibung des Systems
2. Indexierungsprozess
3. Vergleich der Retrieval Modelle
4. Implementierung
5. Zusammenfassung
6. Diskussion
7. Vorführung des fertigen Systems

# Überblick

- Dokumente verwalten und indexieren
- Flexible Suchmaschine
- Verwaltung persönlicher Lesezeichen
- Individuelle Dokument-Empfehlungen

# Konzept



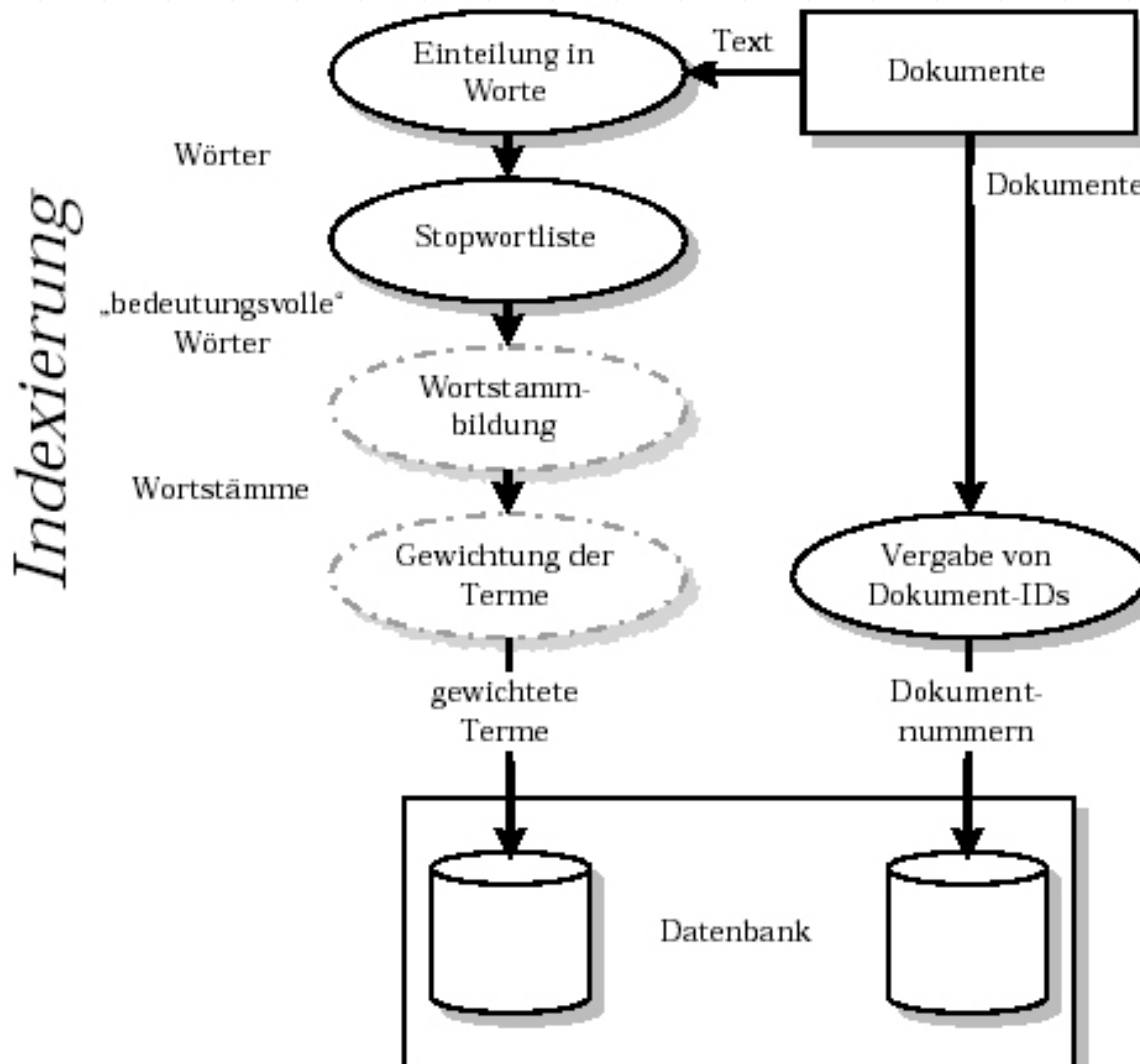
# Webinterface: Suchmaske

Search for "Multimedia OR Bildbearbeitung AND Internet +Server -Email"

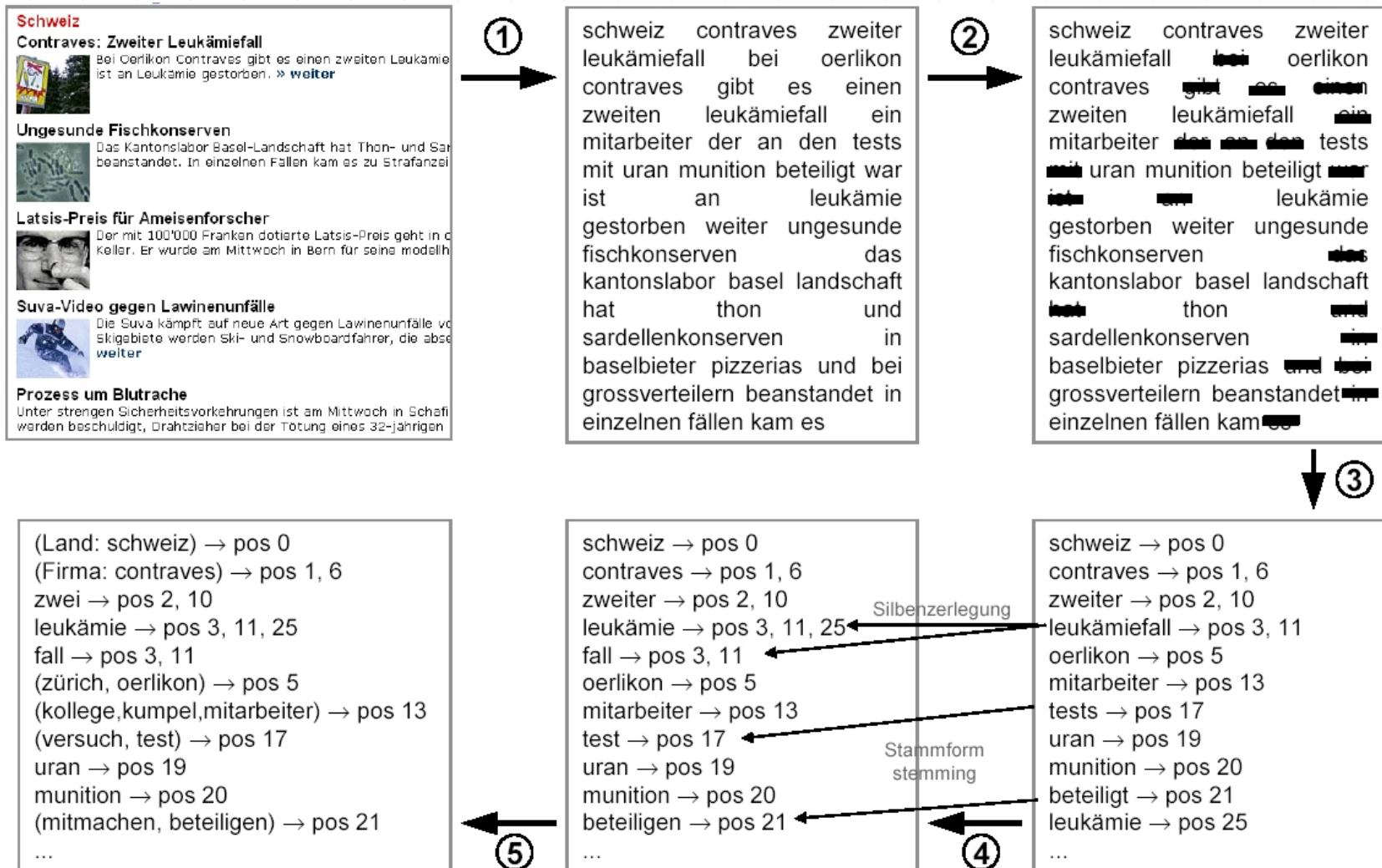
Searchtext	<input type="text" value="Multimedia Bildbearbeitung +Internet +Server -Email"/>	<input type="button" value="Search"/>
Language	<input type="text" value="German"/> ▼	
Filetype	<input type="text" value="[*] All files"/> ▼	
Fileage	<input type="text" value="0"/> [DAYS] to <input type="text" value="1825"/> [DAYS]	
Max results	<input type="text" value="10"/>	
Mark new documents	<input checked="" type="checkbox"/>	
Websearch	<input checked="" type="checkbox"/>	
Favorites	<input type="checkbox"/>	
More Features:	<a href="#">MyPerson-Search</a> <a href="#">Matrikel-Search</a>	

**Note:** Use "+" and "-" to improve your results. [[More help](#)]

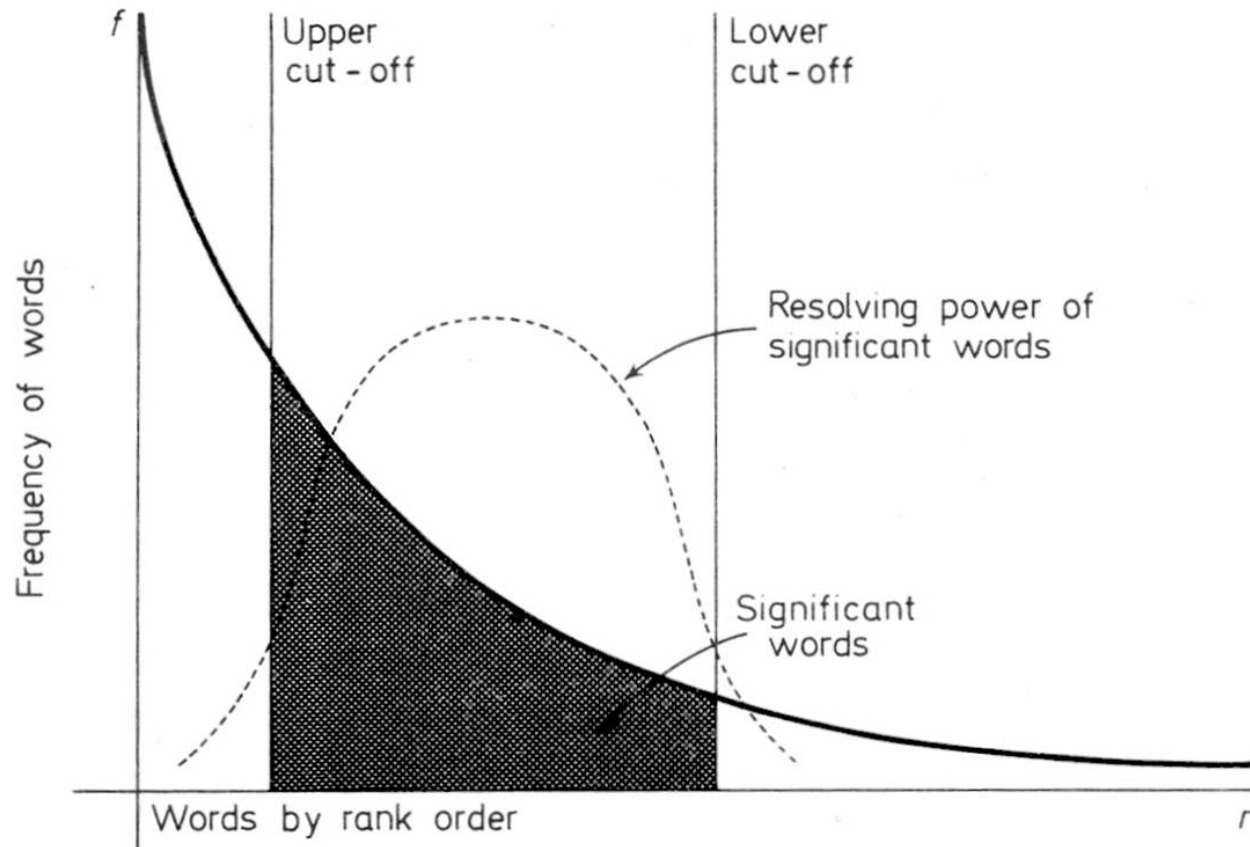
# Ablauf der Indexierung



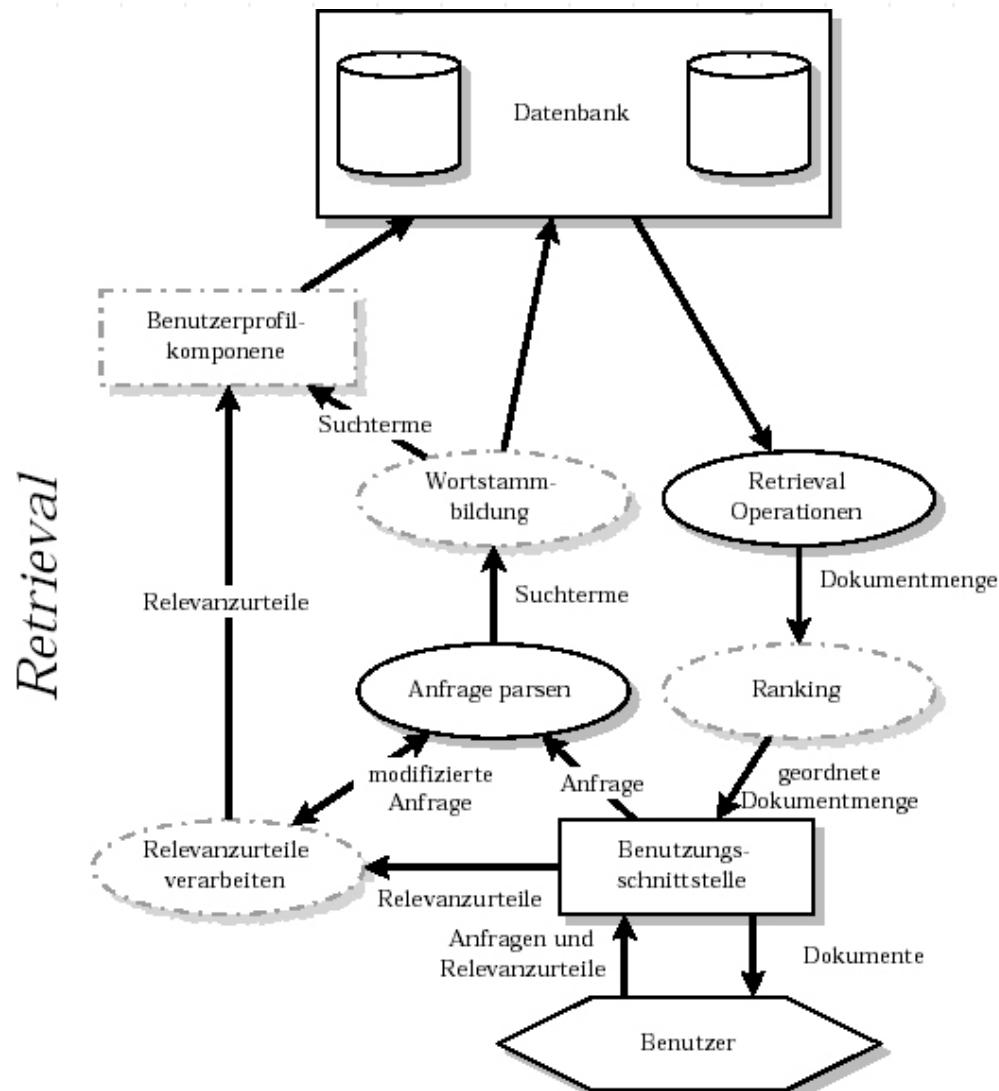
## Beispiel der Indexierung



# Erkennung von Stoppwörtern



# Ablauf des Retrievals



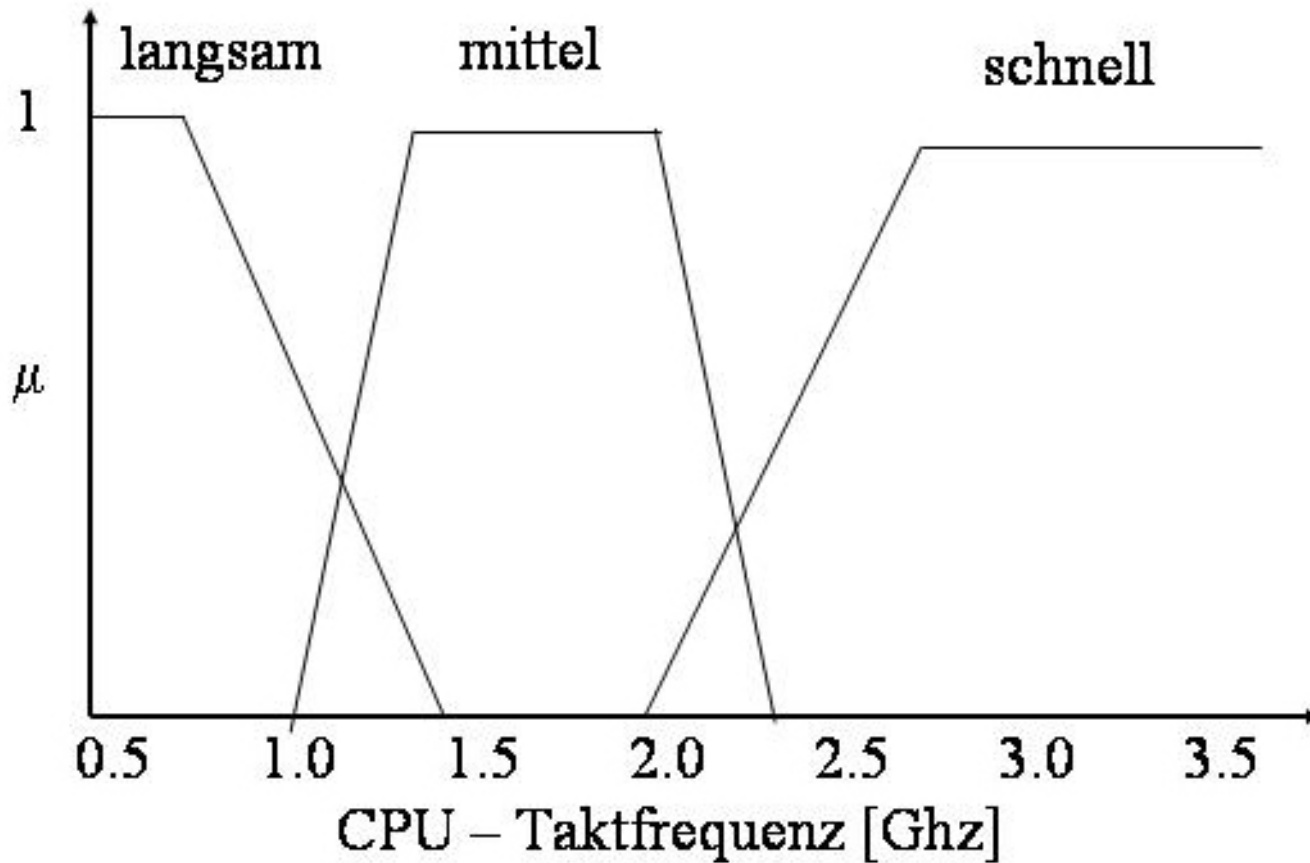
# Retrieval Modelle



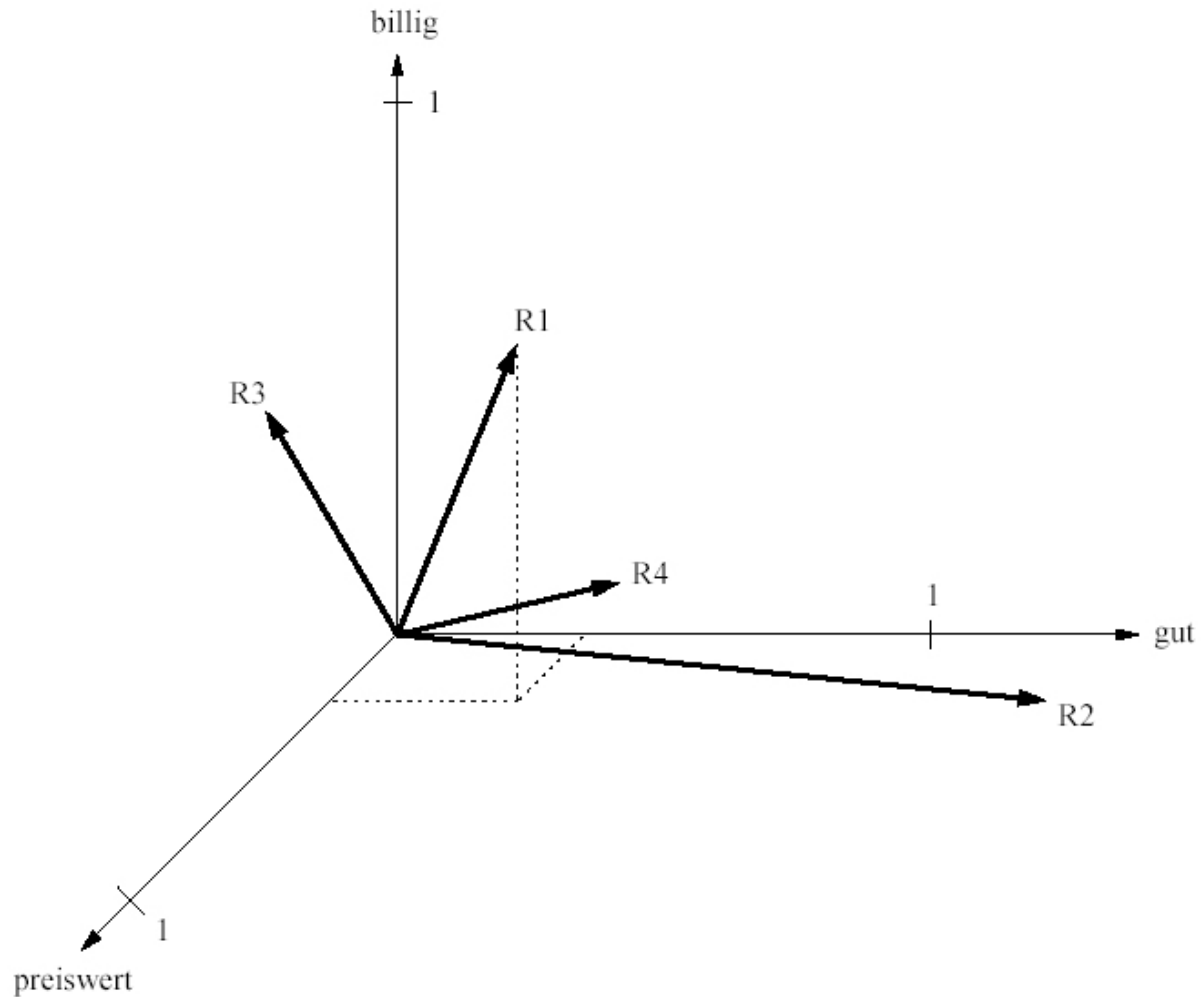
Qualität

- Volltextsuche ( z.B. Boyer-Moore )
- Boolesches Retrieval ( AND, OR, \*, ? )
- Fuzzy Retrieval ( Gewichtung, Häufigkeit )
- Vektorraum Retrieval ( Dokumentvektor )
- Neuronales Netzwerk ( Distanzkarte )

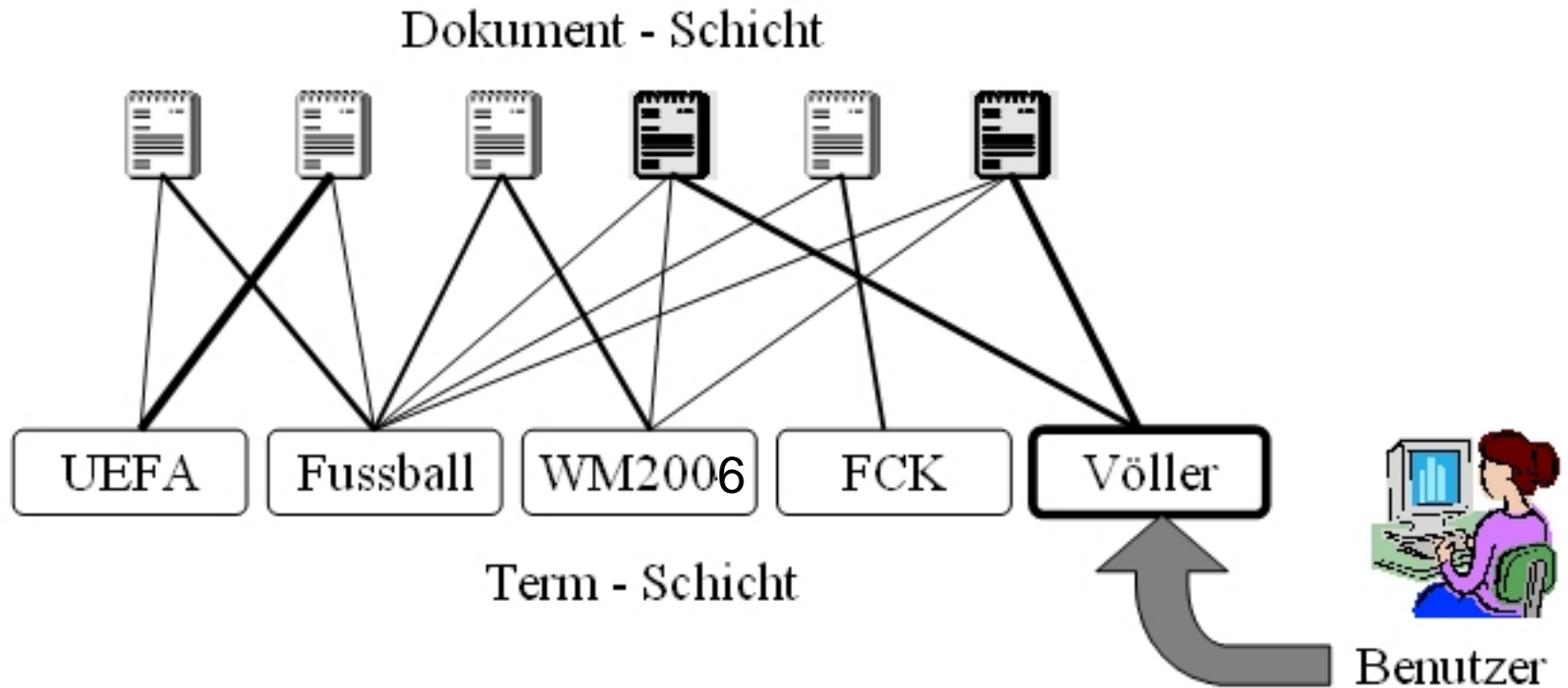
# Fuzzy Modell



# Vektorraum Modell



# Neuronales Netzwerk



## Vergleich

Eigenschaft	Volltextsuche	Boolesche Suche	Fuzzy Suche	Vektorraum Modell	Neuronales Netzwerk
Ranking und Gewichtung	-	-	+	++	++
Natürliche Sprache analysieren	-	-	++	+	+
Vage Ausdrücke interpretieren	-	-	++	-	-
Einfache Implementierung	++	+	O	+	+
Geringe Rechenleistung nötig	++	++	+	++	O

**Bewertung:**

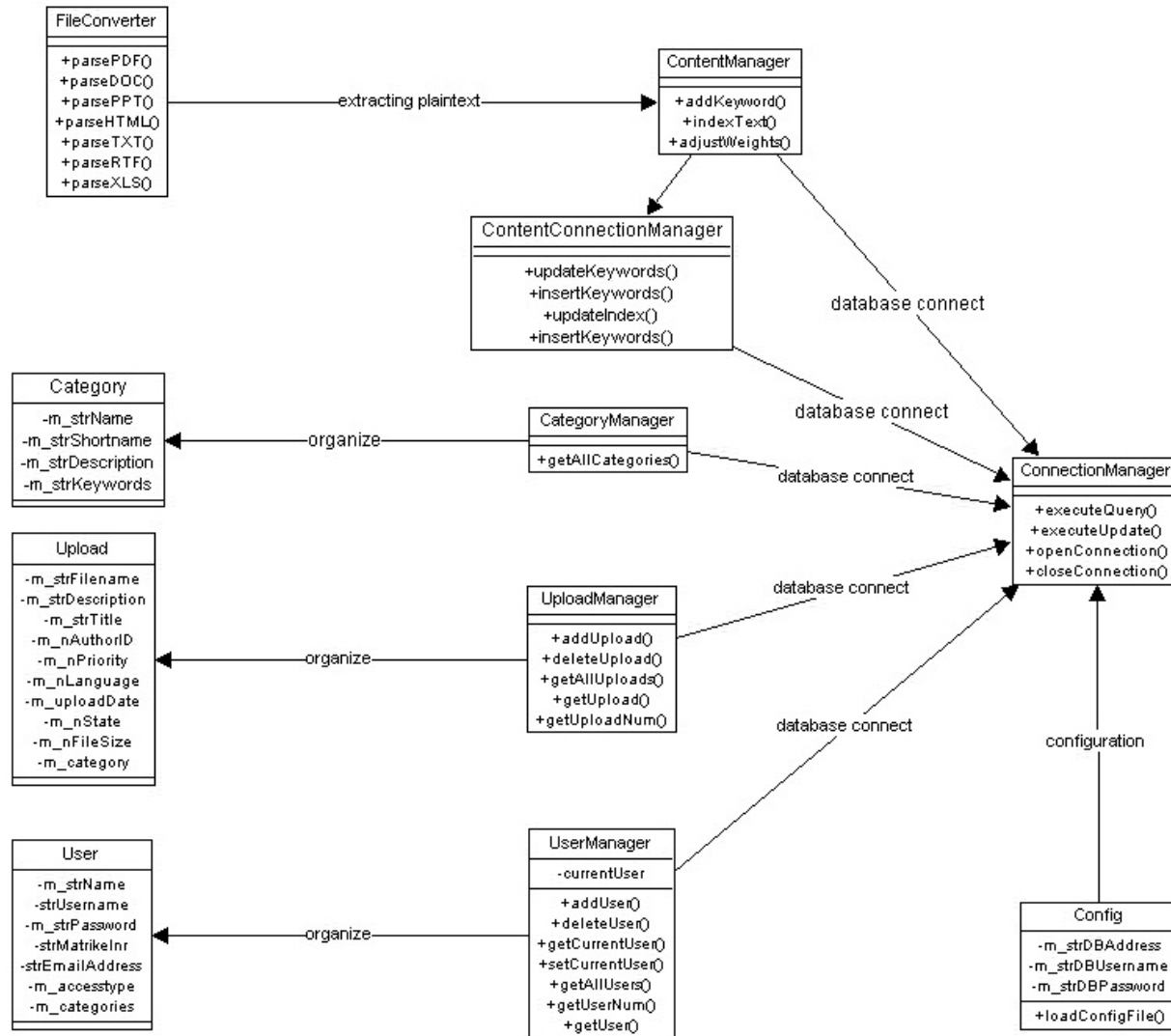
- nicht möglich

O schlecht

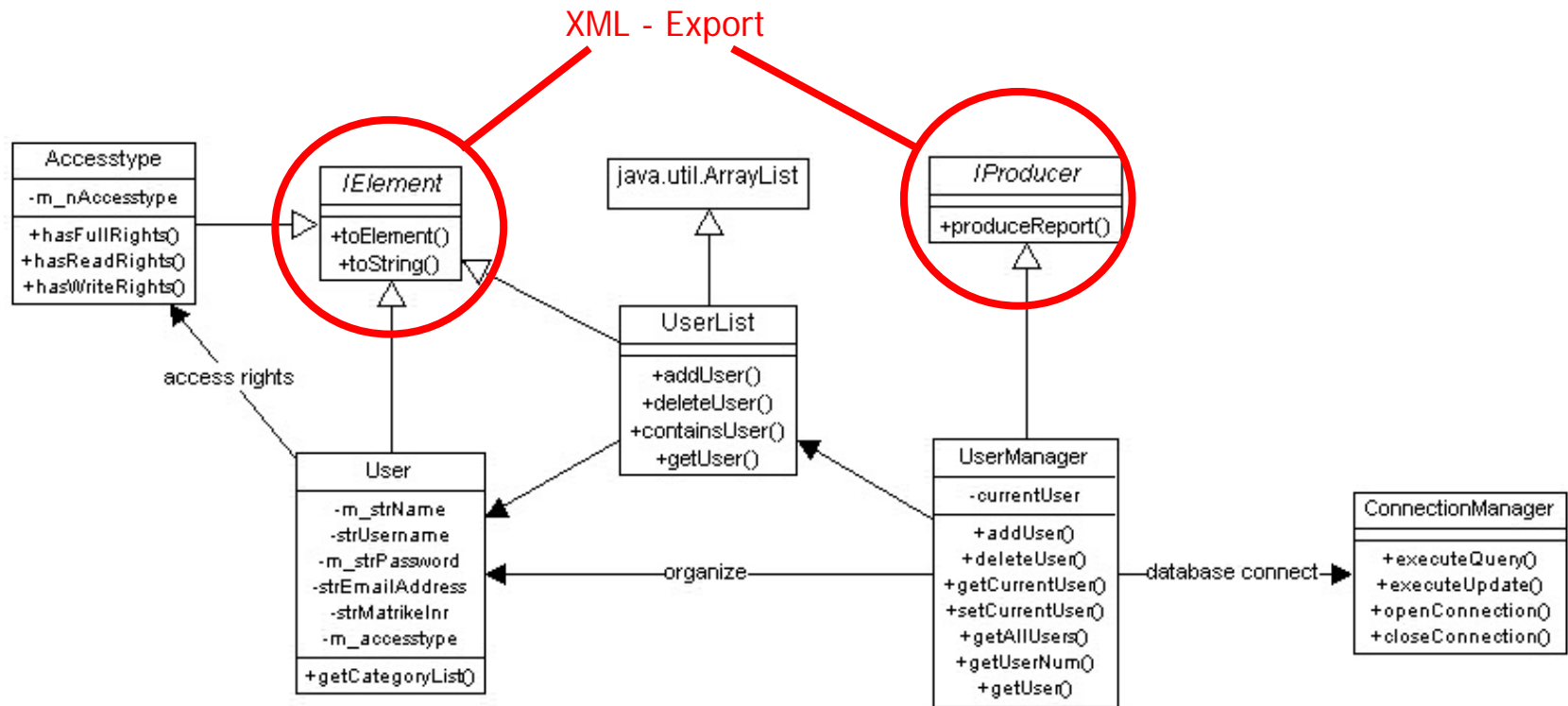
+ gut

++ sehr gut

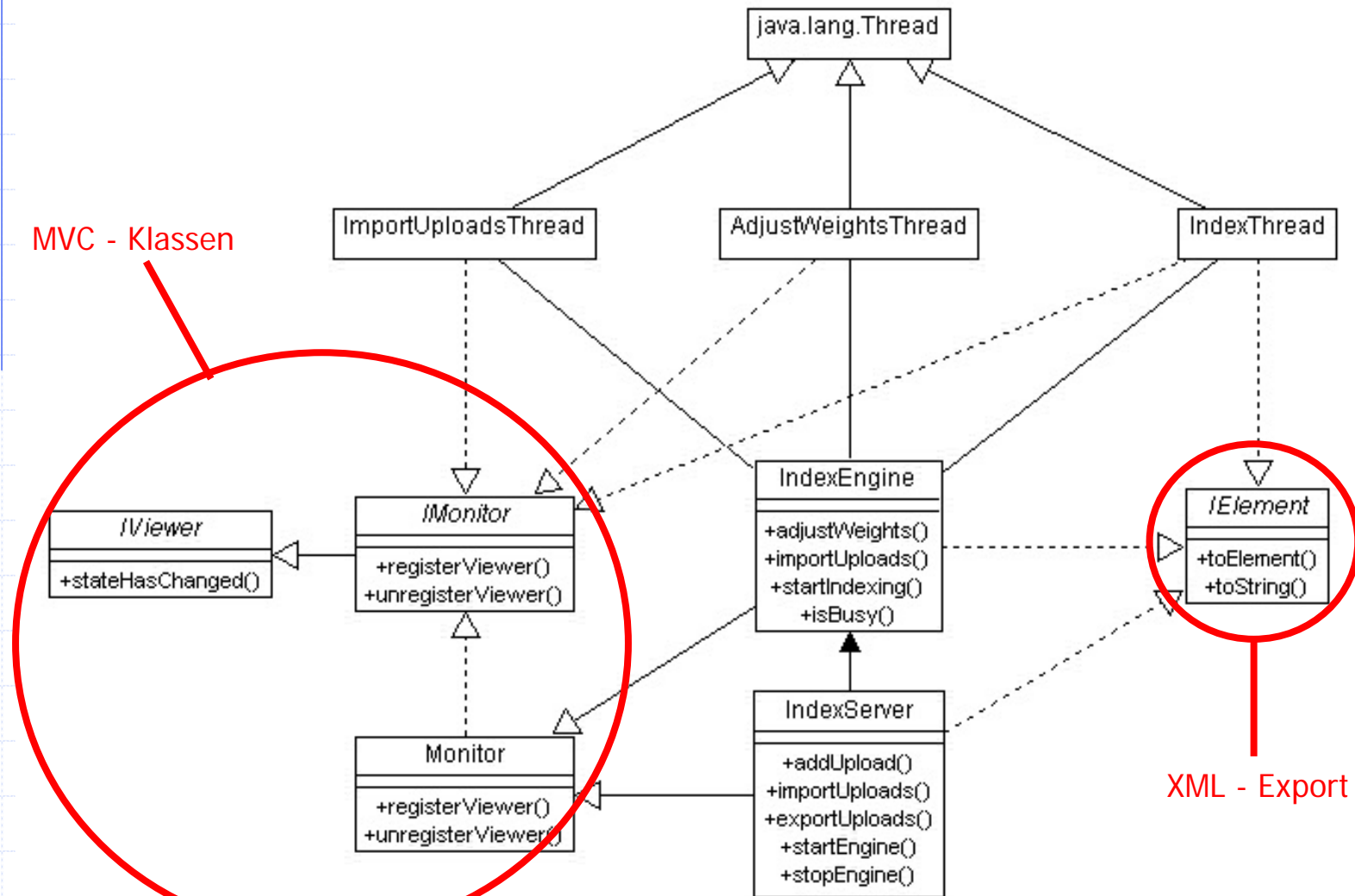
# Framework



# Framework: User



# Indexengine



# Indexthread

```
while ( true )
{
    waitForMemory(); // Wait until enough memory is available, otherwise sleep
    currentUpload = uploadManager.getNextScheduledUpload(); // Get next document
    if ( currentUpload != null )
    {
        try
        {
            contentManager.addUpload ( currentUpload ); // Index document
        }
        catch ( ParsingFailedException e )
        {
            Logger.logWarning ( "File could not be prepared" );
        }
    }

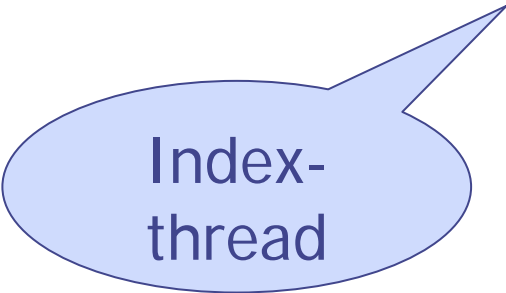
    uploadManager.checkInUpload ( currentUpload ); // Remove from queue
    stateHasChanged(); // Inform viewers about changes
}
else // When queue is empty
{
    sleep ( getEngineSleepTime() ); // Wait some time and check again
}
}
```

# XML - Datenaustausch

```
- <user id="1">
  <name>Jan Löffler</name>
  <username>loeffler</username>
  <password>●●●●●●●●●●</password>
  <email>mail@jlsoft.de</email>
  <matrikelnumber>9921716</matrikelnumber>
  <accesstype id="3">Full</accesstype>
+ <categories>
</user>
```



Benutzer



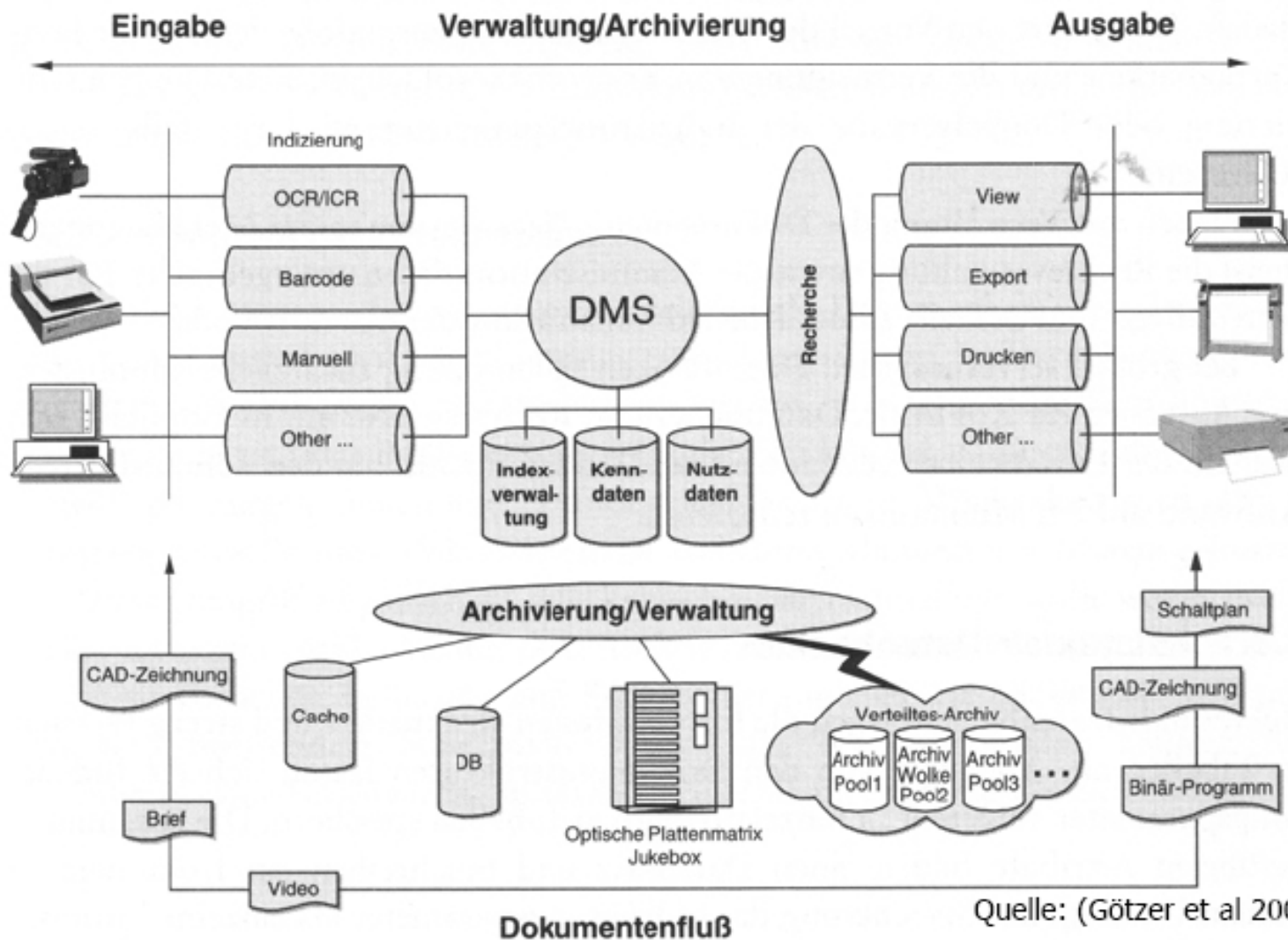
Index-  
thread

```
- <indexengine>
  <busy>true</busy>
  <state>Indexing...</state>
- <indexthread>
  <runtime>00:03:34</runtime>
  <compression>47,00%</compression>
  <prepared>13</prepared>
  <preparedsize>14,69 Mb</preparedsize>
  <speed>4,90 Mb/min</speed>
- <errors>
  <num>1</num>
  <filename>loeffler\APB\apbC1314.pdf</filename>
</errors>
</indexthread>
</indexengine>
```


# Zusammenfassung

- Vektorraum Modell
- Erweiterbarkeit durch Interfaces
- Flexibel durch Framework
- Gute Performance durch relationale DB

# System der Zukunft



# Vielen Dank



...nun zur  
Diskussion