Suche in Metadaten und Volltext

Table of contents

1 Suchfunktionen

MyCoRe unterstützt die kombinierte Suche in den Metadaten von Objekten (Titel, Autor etc.), in den Metadaten von Dateien (Änderungsdatum, Dateigröße etc.) und in den Volltexten von Dateien. Eine MyCoRe-Anwendung sucht dabei nicht direkt in den Daten von Objekten und Dateien, sondern in daraus abgeleiteten Suchfeldern. Die Abbildung von Metadaten auf Suchfelder erfolgt über eine Konfigurationsdatei. So wird z. B. das XML-Element `/metadata/titles/title` auf ein Suchfeld `title` abgebildet.

```
<!-- Search in document metadata -->
<field name="title"          type="text"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="author"         type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="authorID"       type="identifier" source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="authorTitle"    type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="creator"        type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="creatorID"      type="identifier" source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="creatorTitle"   type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="publisher"      type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="publisherID"    type="identifier" source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="publisherTitle" type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="contributor"    type="name"       source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="contributorID"  type="identifier" source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="contributorTitle" type="name"     source="objectMetadata" objects="document" xpath="/mycoreobject/metadata/
<field name="origin"         type="identifier" source="objectCategory" objects="document" xpath="/mycoreobject/metadata/
```

Abbildung 1: Auszug aus der Datei searchfields.xml

Dabei können auch komplexe Suchanfragen mit booleschen Ausdrücken (und/oder/nicht), Platzhaltern und Suchoperatoren gestellt werden. Der Datentyp eines Suchfeldes (ID, Name, Text, Zahl, Datum etc.) bestimmt die bei der Suche einsetzbaren Operatoren (z.B. Phrasensuche, Trunkierung, <, >, ...).

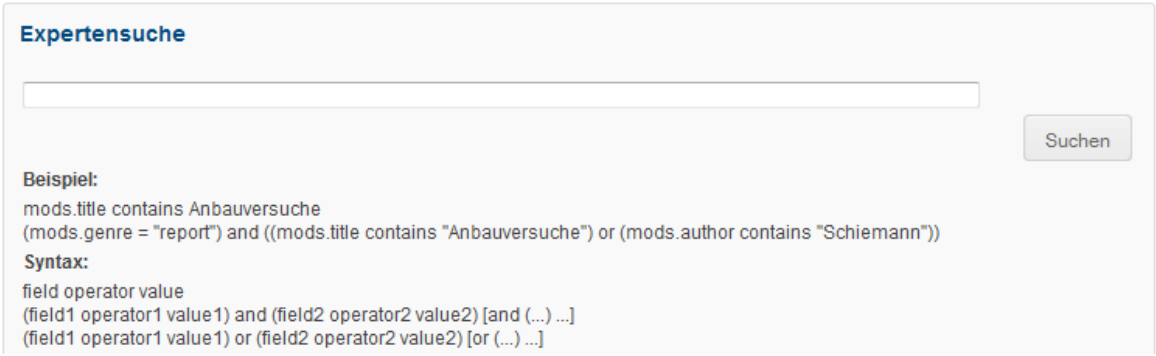

Abbildung 2: Expertensuche mit der MyCoRe-Anfragesprache

Die Suche in MyCoRe ist auf Basis von Apache Lucene (http://lucene.apache.org) realisiert. Obwohl auch andere Implementierungen denkbar sind, ist dies die zur Zeit einzige Implementierung der MyCoRe Suche, die für den Produktionseinsatz sinnvoll ist. Über Lucene durchlaufen indizierte Texte ggf. verschiedene Normalisierungsschritte wie Stammwortreduktion (Stemming) und Umlautnormalisierung.

2 Suchmasken und Trefferlisten

MyCoRe-Suchmasken sind frei konfigurierbar, von einfachen Ein-Feld-Suchformularen über komplexere, qualifiziertere Suchmasken bis hin zu frei formulierbaren Experten-

Abfragen in der MyCoRe-Query-Language (MCRQL). Suchmasken können selbst erstellt oder aus einer Konfiguration von Suchfeldern automatisch generiert werden.

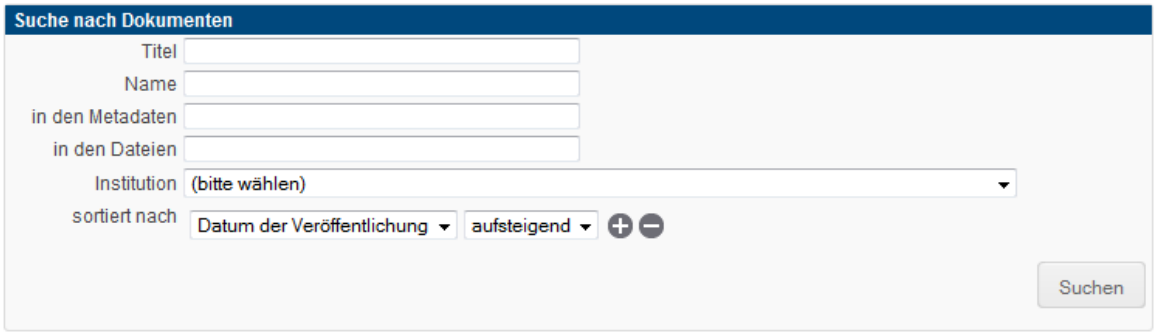

Abbildung 3: Beispiel einer Suchmaske

Trefferlisten sind auf- oder absteigend nach mehreren Feldern beliebig sortierbar. Die ursprüngliche Suche kann angezeigt ("Sie haben gesucht nach: ...") oder noch einmal verfeinert werden (Rückkehr zum Suchformular mit Anpassung der Suchparameter).

Auch eine verteilte Suche über mehrere Server ist möglich und intern über Webservices-Schnittstellen realisiert. Auf diesem Wege können auch Legacy-Systeme mittels einer eigenen Implementierung dieser Schnittstellen auf Java-Basis integriert werden.

3 Datei-Inhalte durchsuchen

Für die Volltextsuche wird der Inhalt von Textdateien über konfigurierbare Filter extrahiert. Derzeit sind Implementierungen für HTML, XML, TXT, OpenOffice Formate und PDF-Dateien enthalten.

MyCoRe kann aus bestimmten Dateitypen zusätzliche Metadaten extrahieren und diese ebenfalls durchsuchbar machen. Derzeit sind Implementierungen für die Extraktion bzw. Suche in EXIF-Metadaten von JPEG-Grafiken (Aufnahmedatum etc.), in ID3-Metadaten von MP3-Audiodateien (Titel, Interpret, Länge etc.) und in Metadaten von PDF-Dokumenten (Seitenzahl, Autor etc.) verfügbar.

MyCoRe kann als Content gespeicherte XML-Dateien qualifiziert durchsuchen. Bei entsprechender Konfiguration könnten z. B. die XML-Strukturen einer `manifest.xml`-Datei eines SCORM-Lernpaketes, oder METS-Metadaten eines Digitalisates durchsucht werden.