Volltext-Funktionen im DFG-Viewer

SRU-/ALTO-Anwendungsprofil

Version 1.0

Redaktion:
Alexander Bigga, Sebastian Meyer, Digitale Bibliothek (Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden), unter Mitarbeit der Techniker-Arbeitsgruppe der DFG-Viewer-Community.

Januar 2016
Zellescher Weg 18
D-01069 Dresden

Inhaltsverzeichnis

1 Einleitung

Der DFG-Viewer unterstützt die Anzeige und Suche in Volltexten. Die dafür nötigen Voraussetzungen an die Digitalisate und die Datenlieferanten werden in diesem Dokument beschrieben.

Zur Suche in den Volltexten wird dem Nutzer ein Suchschlitz angezeigt, sobald eine SRU-Schnittstelle in der übergebenen METS-Datei enthalten ist. Die Ergebnisse werden als kurze Textabschnitte unterhalb des Suchschlitzes auf der Seite angezeigt. Optional werden zusätzlich Bildausschnitte mit den Suchtreffern angezeigt. Wählt der Nutzer ein Ergebnis aus, wird die entsprechende Seite im DFG-Viewer angezeigt und die Suchtreffer im Text hervorgehoben.

Die Anzeige von Volltexten lässt sich in der Werkzeugleiste über die Funktion „Volltexte markieren“ aktivieren. Diese Funktion ist nur aktiv, wenn in der übergebenen METS-Datei ein Verweis auf die Volltexte im ALTO-Format enthalten ist.

Beide Zusatzfunktionen sind für den DFG-Viewer optional und können unabhängig voneinander implementiert werden.

Diese Dokumentation wendet sich daher vornehmlich an Personen und Organisationen, die Metadaten zu digitalisierten Medien erfassen, in verschiedenen Anwendungen zur Verfügung stellen möchten, und an Personen oder Organisationen, die Anwendungen für die Darstellung von digitalisierten Medien entwickeln.

Das vorliegende Anwendungsprofil wird in der Regel gemeinsam mit folgenden Standards angewendet:

- dem METS-Anwendungsprofil[1], das beschreibt, welche Metadaten notwendig sind, um die Struktur digitalisierter Handschriften und Drucke zu beschreiben;
- dem DFG-Viewer Strukturdatenset[2], das beschreibt, welche Strukturtypen in der logischen Struktureinheit der METS-Strukturbeschreibung verwendet werden.

1.1 Hinweise zur Implementierung

Search/Retrieve via URL (SRU)[3] beschreibt einen Standard für Suchanfragen im Internet. Der Suchbegriff wird als *Contextual Query Language* (CQL) per URL übergeben. Die Antwort erfolgt in Form von XML. Der Standard wird von der Library of Congress gepflegt. Grundlage für die Implementierung ist Version 1.2.

Analyzed Layout and Text Object (ALTO)[4] ist ein XML-Schema für die Beschreibung von Layout und Inhalten von Texten, die mittels OCR (*Optical Character Recognition*) erstellt wurden. Das vorliegende Anwendungsprofil setzt mindestens Version 2.0 voraus.

---

[1] http://dfg-viewer.de/profil-der-metadaten/

[2] http://dfg-viewer.de/strukturdatenset/

[3] http://www.loc.gov/standards/sru/

[4] http://www.loc.gov/standards/alto/

Metadaten, die diesem Profil entsprechen, müssen in UTF-8[5] kodiert vorliegen. XML-Daten sind grundsätzlich case-sensitive, die im Anwendungsprofil vorgegebene Groß-/Kleinschreibung von Elementen, Attributen und Werten ist deshalb verpflichtend.

[5] http://tools.ietf.org/html/rfc3629

2 Format der SRU-Anfrage und Antwort

Der folgende Abschnitt beschreibt die in diesem Anwendungsprofil erlaubten SRU-Datenelemente. Dabei folgt die Beschreibung folgendem Aufbau:

SRU-Definition: Gibt die Definition bzw. Beschreibung des Elements oder Unterelements in der SRU Dokumentation[6] wieder.

Kommentar: Enthält profilspezifische Angaben zum Element oder Unterelement.

Wiederholbar: Gibt an, ob ein Element oder Unterelement wiederholbar ist.

Verpflichtungsgrad: Gibt an, ob ein Element oder Unterelement mindestens einmal vorhanden sein muss. Die Verpflichtung kann sich aus einer spezifischen Anforderung des DFG-Viewers und dem allgemeinen METS-Schema ergeben. Es gelten die folgenden Werte:
verpflichtend: das Element muss immer vorhanden sein (wird aber nicht zwangsläufig vom DFG-Viewer interpretiert);
optional: das Element darf vorhanden sein;
konditional: die Verwendung des Elements ist abhängig vom Kontext, in dem es verwendet wird.

Attribute: Nennt die Attribute, die mit einem Element oder Unterelement verwendet werden können oder müssen.

Werte: Nennt die Elementinhalte bzw. deren Wertebereiche, die bei der Verwendung eines bestimmten Elements, Unterelements oder Attributs erlaubt sind.

2.1 Angabe der SRU-Schnittstelle im Digitalisat

Die Angabe der SRU-Schnittstelle im Digitalisat wird im METS-Anwendungsprofil 2.2.1 unter Punkt 2.7.4.3 „SRU-Rechercheschnittstelle – dv:sru“ beschrieben.

Für die hier beschriebene Volltextsuche ist die Angabe verpflichtend.

2.1.1 Beispiel

```
<dv:links>
   <dv:sru>http://digital.slub-dresden.de/sru/356448053</dv:sru>
</dv:links>
```

2.2 Zusammensetzung der SRU Anfrage-URL

Der DFG-Viewer stellt an die SRU-Schnittstelle eine „searchRetrieve“-Anfrage mit dem eingegebenen Suchbegriff. Dazu stellt der DFG-Viewer selbstständig die URL für einen

[6] http://www.loc.gov/standards/sru/

GET-Aufruf zusammen und ruft das Ergebnis von der SRU-Schnittstelle ab. Die im Folgenden aufgeführten Parameter muss die SRU-Schnittstelle mindestens unterstützen. Dies ist auch die Anforderung von SRU 1.2.

2.2.1 Anfrageoperation

Parameter: operation
SRU-Definition: The string „searchRetrieve“
Verpflichtungsgrad: verpflichtend
Wert: searchRetrieve

2.2.2 Version

Parameter: version
SRU-Definition: The version of the request, and a statement by the client that it wants the response to be less than, or preferably equal to, that version.
Verpflichtungsgrad: verpflichtend
Wert: 1.2

2.2.3 Position des ersten Elements in den Suchergebnissen

Parameter: startRecord
SRU-Definition: The position within the sequence of matched records of the first record to be returned. The first position in the sequence is 1. The value supplied MUST be greater than 0. The default value if not supplied is 1.
Verpflichtungsgrad: optional
Wert: 1

2.2.4 Anzahl der Ergebnisse

Parameter: maximumRecords
SRU-Definition: The number of records requested to be returned. The value must be 0 or greater. Default value if not supplied is determined by the server. The server MAY return less than this number of records, for example if there are fewer matching records than requested, but MUST NOT return more than this number of records.
Verpflichtungsgrad: optional
Wert: 10

2.2.5 Suchterm

Parameter: query
SRU-Definition: Contains a query expressed in CQL to be processed by the server.
Verpflichtungsgrad: verpflichtend

2.2.6 Ergebnisschema

Parameter: recordSchema
SRU-Definition: The schema in which the records MUST be returned. The value is the URI identifier for the schema or the short name for it published by the server. The default value if not supplied is determined by the server.
Verpflichtungsgrad: Optional
Wert: dfg-viewer/page

2.2.7 Beispiel

Im Suchfeld wurde eingegeben: „Postzeitung Augsburg“

```
http://digital.slub-
dresden.de/sru/356448053/?operation=searchRetrieve&version=1.2&startRecord=1&maximumReco
rds=10&recordSchema=dfg-viewer/page&query=“Postzeitung%20Augsburg“
```

2.3 Format der SRU Sucherergebnisse

Die SRU-Schnittstelle muss den Funktionsumfang von SRU 1.2 unterstützen. Dabei muss die CQL *(Context Query Language*) mindestens dem Level 0 entsprechen und damit einfache Suchterme unterstützen.

Die einzelnen Suchergebnisse werden im Feld `srw:recordData` zurückgeliefert. Darin eingebettet ist ein DFG-Viewer spezifisches XML.

Der Server antwortet auf die Suchanfrage immer mit einer XML-Struktur, die folgenden Elememente enthält.

2.3.1 Version des SRU-Servers – srw:version

SRU-Definition: The version of the response. This MUST be less than or equal to the version requested by the client.
Kommentar: Enthält die SRU-Version des antwortenden Servers.
Wiederholbar: nein
Verpflichtungsgrad: verpflichtend

2.3.2 Anzahl der Suchergebnisse – srw:numberOfRecords

SRU-Definition: The number of records matched by the query. If the query fails this MUST be 0.

Kommentar: Enthält die Gesamtzahl der Suchergebnisse. Werden keine Ergebnisse gefunden muss dieser Wert 0 sein.

Wiederholbar: nein

Verpflichtungsgrad: verpflichtend

2.3.3 Wiederholung der Suchanfrage – srw:echoedSearchRetrieveRequest

SRU-Definition: The request parameters echoed back to the client in a simple XML form.

Kommentar: Enthält die gestellte SRU-Anfrage. Z.B. startRecord, maximumRecords und query. Siehe Abschnitt 2.2.

Wiederholbar: nein

Verpflichtungsgrad: optional

2.3.4 Suchergebnisse im DFG-Viewer spezifischer Namensraum

Der Inhalt eines SRU-Suchergebnisses wird im Element `srw:recordData` zurückgeliefert. Dieses Element bietet die Möglichkeit ein anwendungsspezifisches XML mit eigenem Namensraum zurückzuliefern. Der XML-Namensraum für den DFG-Viewer muss wie folgt deklarieren werden:

```
@xmlns:dv="http://dfg-viewer.de/"
```

2.3.4.1 Beschreibung der Seite – dv:page

Kommentar: Beschreibt die physischen Dimensionen des Bildes, welches den Suchtreffer enthält.

Wiederholbar: nein

Verpflichtungsgrad: verpflichtend

Attribute: Die Angabe der folgenden Attribute ist verpflichtend:

- **id**: @ID der das Bild repräsentierenden physischen Struktur (METS-Anwendungsprofil 2.2.1)
- **width**: die Breite des Bildes in Pixel
- **height**: die Höhe des Bildes in Pixel

2.3.4.2 Referenz auf die METS-Datei – dv:page/dv:parent

Kommentar: Beschreibt die Seite des Digitalisats, die den Suchtreffer enthält.

Wiederholbar: nein

Verpflichtungsgrad: verpflichtend

Attribute: Die Angabe der folgenden Attribute ist verpflichtend:
- **id**: @ID der die Seite repräsentierenden logischen Struktur (METS-Anwendungsprofil 2.1.1)
- **url**: die URL der METS-Datei des Digitalisats, die das Suchergebnis auf der angegebenen Seite enthält.

Wert: Enthält den Haupttitel des Digitalisats (mods:titleInfo/mods:title)

2.3.4.3 Paginierung der Seite – dv:page/dv:pagination

Kommentar: Enthält die Paginierung der Einzelseite

Wiederholbar: nein

Verpflichtungsgrad: optional

2.3.4.4 Suchtreffer Text- und Bildausschnitt – dv:page/dv:fulltexthit

Kommentar: Enthält die Geometrie des gefundenen Wortes. Optional kann auch ein Bildausschnitt mit dem markierten Wort angegeben werden.

Wiederholbar: nein

Verpflichtungsgrad: verpflichtend

Attribute:
- **x1, y1, x2, y2**: Pixelkoordinaten des Suchtreffers in Bezug auf die in `dv:page` angegebenen Bilddimensionen (verpflichtend)
- **preview**: URL des Bildausschnitts mit Suchtreffer (optional)

2.3.4.5 Volltext Text – dv:page/dv:fulltexthit/dv:span

Kommentar: Enthält einen Textausschnitt vor und nach dem gefundenen Suchtreffer

Wiederholbar: nein

Verpflichtungsgrad: verpflichtend

Attribute:
- **class**: optionale CSS-Klasse für die Darstellung. Der DFG-Viewer unterstützt die Klasse „highlight“.

2.3.5 Beispiele

Wiederholung der Suchanfrage in der Antwort:

```
<srw:echoedSearchRetrieveRequest>
   <srw:version>1.2</srw:version>
   <srw:startRecord>1</srw:startRecord>
   <srw:maximumRecords>10</srw:maximumRecords>
   <srw:query>Postzeitung Augsburg</srw:query>
```

```
   <srw:recordSchema>dfg-viewer/page</srw:recordSchema>
</srw:echoedSearchRetrieveRequest>
```

Einzelnes Suchergebnis mit Vorschautext und Vorschaubild:

```
<dv:page id="phys30439" width="2215" height="3076"
url="http://visuallibrary.net/download/webcache/0/30439" xmlns:dv="http://dfg-
viewer.de/">
   <dv:parent url="http://visuallibrary.net/mets/vd/id/228591"
   id="log228591">20.8.1854 (No. 34)</dv:parent>
   <dv:pagination>268</dv:pagination>
   <dv:fulltexthit x1="1561" y1="1229" x2="1669" y2="1275"
   preview="http://visuallibrary.net/search/pagecrop?id=30439&amp;term=post">
      <dv:span>vorzüglich berufen. Als daher der damit bethcilte </dv:span>
      <dv:span class="highlight">Posten</dv:span>
      <dv:span> eines Gubernialraihes zu Innsbruck zu besetzen kam,
      erfolgte</dv:span>
   </dv:fulltexthit>
</dv:page>
```

3 Volltexte im ALTO-Format

3.1 Einbindung der ALTO-Volltexte in die METS-Datei

Der DFG-Viewer kann den Volltext eines Digitalisats anzeigen. Dazu muss die METS-Datei eine Dateigruppe `mets:fileGrp` mit dem Attribut `FULLTEXT` enthalten (METS-Anwendungsprofil 2.2, „2.4.2.1 Dateigruppen – mets:fileGrp“[7]). Die Volltexte müssen im ALTO-Format[8] mit Wortkoordinaten vorliegen. Das ALTO-Format muss mindestens der Schema-Version 2.0 entsprechen.

Findet der DFG-Viewer eine solche Dateigruppe für die aktuelle Seite, lässt sich die Funktion „Volltexte markieren“ in der Werkzeugleiste aktivieren. Die Volltexte werden dann in einem separaten Bereich angezeigt.

3.1.1 Beispiel:

Einbinden der Volltexte in die METS-Datei:

```
<mets:fileGrp USE="FULLTEXT" >
   <mets:file ID="FILE_0001_FULLTEXT" MIMETYPE="text/xml" >
      <mets:FLocat LOCTYPE="URL" xlink:href="http://digital.slub-
      dresden.de/fileadmin/data/359023940/359023940_ocr/00000001.xml"
      ></mets:FLocat>
   </mets:file>
</mets:fileGrp>

Volltexte in ALTO mit Wortkoordinaten:
/alto/Layout/Page/PrintSpace/TextBlock/TextLine
<TextLine HEIGHT="28" WIDTH="418" VPOS="351" HPOS="98">
   <String CONTENT="den" HEIGHT="28" WIDTH="47" VPOS="351" HPOS="98"/>
   <SP WIDTH="14" VPOS="362" HPOS="146"/>
   <String CONTENT="und" HEIGHT="28" WIDTH="57" VPOS="351" HPOS="161"/>
   <SP WIDTH="13" VPOS="351" HPOS="219"/>
   <String CONTENT="im" HEIGHT="27" WIDTH="38" VPOS="352" HPOS="233"/>
   <SP WIDTH="13" VPOS="352" HPOS="272"/>
   <String CONTENT="Lande" HEIGHT="28" WIDTH="91" VPOS="351" HPOS="286"/>
   <SP WIDTH="13" VPOS="352" HPOS="378"/>
   <String CONTENT="Sachsen." HEIGHT="28" WIDTH="124" VPOS="351" HPOS="392"/>
</TextLine>
```

[7] http://dfg-viewer.de/profil-der-metadaten/

[8] http://www.loc.gov/standards/alto/