



Herzlich Willkommen

Betrachtung von Grenzfällen in der OCR und Suche
Michael Kubina (*Digitale Dienste*)

Kitodo Praxistreffen 2024
Hessisches Landesarchiv | Marburg, 18.11.2024

ÜBERSICHT

- ➔ Multiscriptualität / -lingualität
... und wie wir mit zeichengenauer OCR umgehen
- ➔ Zeitbasierte Suche
... und welche Dimensionen uns noch fehlen
- ➔ Suche im Dokument
... und warum viel Erfassungsaufwand ungenutzt bleibt



Multiscriptualität / -lingualität

- Immer größeres Interesse von Forschenden und Lehrenden nach Unterstützung von Multiscriptualität / -lingualität
- Fachverband DHd AG Multilingual DH
 - <https://dig-hum.de/ag-multilingual-dh>
- DARIAH-EU Multilingual DH Working Group
 - <https://multilingual.hypotheses.org/>
- SUB-HH Referat Digitale Forschungsdienste
- Centre for the Study of Manuscript Cultures
- Exzellenzcluster: Understanding Written Artefacts
- Digitalisierte Bestände inkl. externe Einrichtungen
 - Asien-Afrika-Institut
 - Institut für die Geschichte der deutschen Juden
 - Museum am Rothenbaum, Kulturen und Künste der Welt
 - eigene historische Bestände
 - perspektivisch auch Handschriften

Kitodo.Production 2 Regelsatz

```
<select1 tns:ref="DocLanguage">
  <item tns:selected="false">
    <label>Arabisch</label>
    <value>ara</value>
  </item>
  <item tns:selected="false">
    <label>Aramäisch</label>
    <value>arc</value>
  </item>
  <item tns:selected="false">
    <label>Armenisch</label>
    <value>arm</value>
  </item>
  <item tns:selected="false">
    <label>Chinesisch</label>
    <value>chi</value>
  </item>
  <item tns:selected="false">
    <label>Altkirchenslawisch</label>
    <value>chu</value>
  </item>
  <item tns:selected="false">
    <label>Tschechisch</label>
    <value>cze</value>
  </item>
</select1>
```

```
<select1 tns:ref="Font">
  <item tns:selected="false">
    <label>Antiqua</label>
    <value>antiqua</value>
  </item>
  <item tns:selected="false">
    <label>Fraktur</label>
    <value>fraktur</value>
  </item>
  <item tns:selected="true">
    <label>Ohne_OCR</label>
    <value>Ohne_OCR</value>
  </item>
</select1>
```

Kitodo.Production 3 Regelsatz

```
<key id="language">
  <label>Language (Group)</label>
  <label lang="de">Sprache (Gruppe)</label>
  <key id="languageTerm">
    <label>Language</label>
    <label lang="de">Sprache</label>
  </key>
  <key id="scriptTerm">
    <label>Script</label>
    <label lang="de">Schrift</label>
    <option value="Latn"><label>a) Lateinisch (Antiqua)</label></option>
    <option value="Latf"><label>b) Lateinisch (Fraktur)</label></option>
    <option value="Arab"><label>c) Arabisch</label></option>
    <option value="Grek"><label>d) Griechisch</label></option>
    <option value="Hebr"><label>e) Hebräisch</label></option>
    <option value="Cyrl"><label>f) Kyrillisch</label></option>
    <option value="automatic"><label>g) Sonstige Schrift / Automatisch</label></option>
    <option value="handwriting"><label>h) Handschriftlich</label></option>
    <option value="musical_notation"><label>i) Musiknotation</label></option>
    <option value="Zxxx"><label>j) Nicht geschriebenes Dokument</label></option>
  </key>
</key>
```

Metadatenbearbeitung

▼ Sprache (Gruppe):



Sprache:

ger



Schrift:

- a) Lateinisch (Antiqua)
- b) Lateinisch (Fraktur)
- c) Arabisch
- d) Griechisch
- e) Hebräisch
- f) Kyrillisch
- g) Sonstige Schrift / Automatisch
- h) Handschriftlich
- i) Musiknotation
- j) Nicht geschriebenes Dokument



XML-Schema

meta.xml

```
<kitodo:metadataGroup name="language">  
  <kitodo:metadata name="scriptTerm">Latf</kitodo:metadata>  
  <kitodo:metadata name="languageTerm">ger</kitodo:metadata>  
  <kitodo:metadata name="scriptTerm">Latn</kitodo:metadata>  
</kitodo:metadataGroup>
```

MODS-Schema⁽¹⁾

```
*****  
* Top Level Element <language> *  
*****  
-->  
<xs:element name="language" type="languageDefinition"/>  
<!-- -->  
<xs:complexType name="languageDefinition">  
  <xs:sequence>  
    <xs:element ref="languageTerm" maxOccurs="unbounded"/>  
    <xs:element ref="scriptTerm" minOccurs="0" maxOccurs="unbounded"/>  
  </xs:sequence>
```

METS/MODS

```
<mods:language>  
  <mods:languageTerm type="code" authority="iso639-2b">ger</mods:languageTerm>  
  <mods:scriptTerm type="code" authority="iso15924">Latf</mods:scriptTerm>  
  <mods:scriptTerm type="code" authority="iso15924">Latn</mods:scriptTerm>  
</mods:language>
```

DDB-Anwendungsprofil⁽²⁾

Verwendung

Für Textdokumente ist die Angabe der Sprache verpflichtend. Wird in dem Dokument mehr als eine Sprache verwendet, wird mods:language wiederholt.

Verpflichtende Unterelemente in mods:language

```
<mods:languageTerm>
```

Enthält den ISO 639-2b Code der in dem Dokument verwendeten Sprache.

DFG-Anwendungsprofil⁽³⁾

MODS Anwendungsprofil für digitalisierte Medien 2.3.1

2.5.2.2 Schrift – mods:scriptTerm

1. <https://www.loc.gov/standards/mods/mods.xsd>
2. <https://wiki.deutsche-digitale-bibliothek.de/pages/viewpage.action?pageId=69124887>
3. https://dfg-viewer.de/fileadmin/groups/dfigviewer/MODS-Anwendungsprofil_2.3.1.pdf

Use Case: Automatic Text Recognition

```
# wahl des schriftmodells durch die in kitodo ausgewaehlte schrift
case $2 in
  "Latn")
    TESSMODEL+="script/Latin"
    ;;
  "Latf")
    TESSMODEL+="script/Fraktur"
    ;;
  "Arab")
    TESSMODEL+="script/Arabic"
    ;;
  "Grek")
    # spezialmodel (alt-)griechisch
    TESSMODELPRIO+="external/grc_hist"
    # fallback auf griechisch
    TESSMODEL+="script/Greek"
    ;;
  "Hebr")
    TESSMODEL+="script/Hebrew"
    ;;
  "Cyril")
    TESSMODEL+="script/Cyrillic"
    ;;
esac
# schrift durch automatische auswahl ermitteln, falls in kitodo ausgewaehlt
if [[ $2 == "automatic" ]]; then
  case $1 in
    #Afrikaans
    "afr")
      TESSMODEL+="script/Latin+script/Arabic"
      ;;
    #Albanisch
    "alb")
      TESSMODEL+="script/Latin"
      ;;
    #Amharisch
    "amh")
      TESSMODEL+="script/Ethiopic"
      ;;
    #Arabisch
    "ara")
      TESSMODEL+="script/Arabic"
      ;;
```

```
<language type="awa" territories="IN" alt="secondary"/>
<language type="ay" scripts="Latn" territories="BO"/>
<language type="az" scripts="Arab Cyril Latn" territories="AZ"/>
<language type="az" territories="IQ IQ RO" alt="secondary"/>
<language type="ba" scripts="Cyr1"/>
<language type="ba" territories="RU" alt="secondary"/>
<language type="bal" scripts="Arab"/>
<language type="bal" scripts="Latn" territories="IR PK" alt="secondary"/>
<language type="ban" scripts="Latn"/>
<language type="ban" scripts="Bali" territories="ID" alt="secondary"/>
```



```
<language type="az" scripts="Arab Cyril Latn">
```

<https://github.com/unicode-org/cldr/blob/main/common/supplemental/supplementalData.xml>

Problem: Sprach- & Schriftmix

«Артсанія», можно читать и «Арманія», то-есть, люди изъ Арма или племя Арма. Опушение буквы б въ началѣ, конечно, странно, но объясняется довольно просто: за арабскимъ глаголомъ о—, «самма», называть, имя обыкновенно слѣдуетъ съ предлогомъ л, б, иногда же оно опускается. Первоначально въ источн икъ стояло: «племя это называлось: **بآرمانجی** «В аг ш апу»), что произносилось: «Біарманія», то-есть, Біармійцы, люди изъ Біармы; авторъ или, можетъ быть, переписчикъ, полагалъ, что начальное л, б, не принадлежитъ къ имени, но есть просто предлогъ, который можно и опустить; они опустили его и такимъ образомъ изъ **بآرمانجی** сдѣлали т. е. изъ Біармійцевъ сдѣлали Армійцевъ я). Что эта догадка справедлива, доказательствомъ тому, по моему мнѣнію, можетъ служить слѣдующее: имя столицы означеннаго племени пишется у различныхъ арабскихъ писателей: «Арба», **آربا**, «Арта» или **آرثا**, «Артса»; эта группа буквъ, какъ замѣчено, легко можетъ происходить изъ ЦД «Арма». У эль-Балхи же это имя два раза пишется, «Абарка» и третій разъ **آربا**, «Ар<i>i>а»; і, ф и і, к, собственно, не есть варианты, такъ какъ этихъ буквъ почти нельзя отличать другъ отъ друга въ рукописямъ; послѣднія же четыре буквы, **آربا**, арка или **آرثا**, ар<i>i>а, какъ и **آرثا**, арба,

«Артсанія», можно читать и «Арманія», то-есть, люди изъ Арма или племя Арма. Опушение буквы б въ началѣ, конечно, странно, но объясняется довольно просто: за арабскимъ глаголомъ **سمى**, «самма», называть, имя обыкновенно слѣдуетъ съ предлогомъ **ب**, б, иногда же оно опускается. Первоначально въ источн икъ стояло: «племя это называлось: **بآرمانجی** «Bārmānjī»), что произносилось: «Біарманія», то-есть, Біармійцы, люди изъ Біармы; авторъ или, можетъ быть, переписчикъ, полагалъ, что начальное **ب**, б, не принадлежитъ къ имени, но есть просто предлогъ, который можно и опустить; они и опустили его и такимъ образомъ изъ **بآرمانجی** сдѣлали **آرمانجی**, т. е. изъ Біармійцевъ сдѣлали Армійцевъ я). Что эта догадка справедлива, доказательствомъ тому, по моему мнѣнію, можетъ служить слѣдующее: имя столицы означеннаго племени пишется у различныхъ арабскихъ писателей: **آربا**, «Арба», **آرثا**, «Арта» или **آرثا**, «Артса»; эта группа буквъ, какъ замѣчено, легко можетъ происходить изъ **آرما**, «Арма». У эль-Балхи же это имя два раза пишется **آبارقا**, «Абáркá» и третій разъ **آرثا**, «Арфа»;

Problem: Sprach- & Schriftmix

وابر ها الوقوف على المولدات الثلاثة وما ابلن فيها من قوى م اكب
 ابر وهي المعرب عبالخواش عند القاص ب ولا لحون لما طة ولا
 حبة ولا حاجة الى كتف ن الاو اذل ا نم همزاج بعضها عن بس لجكل 9
 ويتولخي ب حرارة عب ذلك نيل الدخات كل يتمان ،لقوى الك ملا
 النامة) او يتولخي ما حرارة طبيعية فذلك قم المطعومات ا وتلكان لا ،<
 تتغذي ب ولا يمتعان الايا لنخي الانمالية والجواب) والجل المائة بنير درجات 12
 احن انواع هذا الحر العمل ا
 واطم ها اخي ان صن الحر ما هو متقاد وطه ما هو حيلي فن المتاد
 :8: 8: 116؟0: ع1110|| 8: 80 £ ا 0 طنا ا 01. 0 ¥ - بيته : فقد 2 اء.
 وضعوه 1
 ها09: 0 6160: 0: 8* 6008: 0.9. 0: ج913: 1.98: 1110(1 ٢
 068: ع913: 0(1 30: 0: 113٢6: 36
 لل 60: 60: 260 لا 00* 6 لل 1 جلا: 0: 6: 8: 111 00: 8: 111 1٢(11
 ؛| 1١0| : 1(111 11 01: 111 8: 111 10٢: 1٢
 هبع: ٥٥٥٥ ا| < > ا 01. 0 3 ! : ٤ لك : ٢ آء 00 هلا: -
 لل 3^068: 1 98: لا 19: 8ا ال 6ل

فذلك قبيل الدخات الى يستعان بالسوى النامة
 ما حرارة طبيعية فذلك قسم المطعومات ، وتلكان لا
 بالنفس الانسانية والحيوانية ، والحيل المسماة بنيرانجات 2
 العملي
 السحر ما هو مستفاد ومنه ما هو حيلي فمن المستفاد

O طلسمًا | V om. C تقيه : فقد 2 | C₁ وضعوه 1
 aequare in faciendo imagines quod virtus imaginis t
 horarum et temporum ad habendam propriam cons
 rebus, quibus imagines funduntur. Pic. | 3 |
 ratur | 6 Et ex hoc puncto quod verbum habet in s
 erit major fortitudo quando plures fortitudines ad in
 est completa virtus nigromantiae et haec est Theor
 ejus manerioi sit, nec qualiter adjungatur virtus pra

Problem: Zeichengenaue OCR

ä^e

ü^e

ö^e

ı



Aktuelle schema.xml

Analyzer	Abkürzung	Index	Query
OCR Character Filter	OCF	Gefchäfte	
HTML Special Character Filter	HTMLSCF	Gefchäfte	
Whitespace Tokenizer	WT	Gefchäfte	Geschäft
SynonymGraphFilter	SGF		Geschäft
FlattengraphFilter	FGF	Gefchäfte	
Stop(word) Filter	SF	Gefchäfte	Geschäft
WordDelimiterGraphFilter	WDGF	Gefchäfte	Geschäft
ReverseWildcardSearchFilter	RWF	Gefchäfte	
LowercaseFilter	LCF	gefchäfte	geschäft
TrimFilter	TF	gefchäfte	geschäft
SnowballFilter	SF	gefchäft	geschafft
RemoveDuplicateTokenFilter	RDTF	gefchäft	geschafft

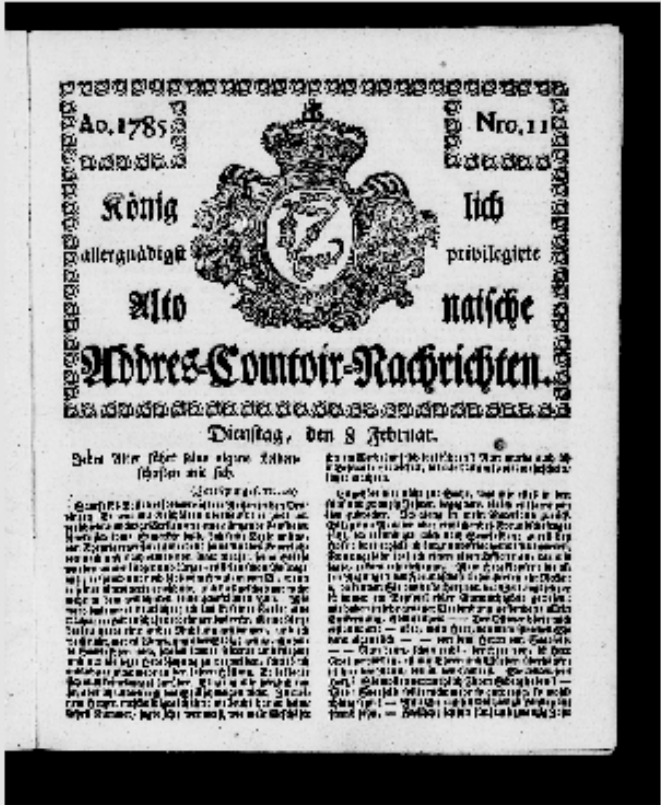
Lösung? Character Folding?

Analyzer	Abkürzung	Index	Query
OCR Character Filter	OCF	Gefchäfte	
HTML Special Character Filter	HTMLSCF	Gefchäfte	
Whitespace Tokenizer	WT	Gefchäfte	Geschäft
AsciiFoldingFilter	ASCIIF	Geschäfte	Geschäft
SynonymGraphFilter	SGF	Geschäfte	Geschäft
FlattengraphFilter	FGF	Geschäfte	
Stop(word) Filter	SF	Geschäfte	Geschäft
WordDelimiterGraphFilter	WDGF	Geschäfte	Geschäft
ReverseWildcardSearchFilter	RWF	Geschäfte	
LowercaseFilter	LCF	geschäfte	geschäft
TrimFilter	TF	geschäfte	geschäft
SnowballFilter	SF	geschäfte	geschäft
RemoveDuplicateTokenFilter	RDTF	geschäfte	geschäft

Lösung? Character Mapping?

Analyzer	Abkürzung	Index	Query
OCR Character Filter	OCF	Gefchäfte	
HTML Special Character Filter	HTMLSCF	Gefchäfte	
MappingCharacterFilter	MCF	Gefchäfte	Geschäft
Whitespace Tokenizer	WT	Gefchäfte	Geschäft
AsciiFoldingFilter	ASCIIF	Geschäfte	Geschäft
SynonymGraphFilter	SGF	Geschäfte	Geschäft
FlattengraphFilter	FGF	Geschäfte	
Stop(word) Filter	SF	Geschäfte	Geschäft
WordDelimiterGraphFilter	WDGF	Geschäfte	Geschäft
ReverseWildcardSearchFilter	RWF	Geschäfte	
LowercaseFilter	LCF	geschäfte	geschäft
TrimFilter	TF	geschäfte	geschäft
SnowballFilter	SF	geschäft	geschäft
RemoveDuplicateTokenFilter	RDTF	geschäft	geschäft

Noch nicht ganz ...



Seite 1
Titel Königlich privilegirte Altonaer Adreß-Comtoir-Nachrichten

[...] ihn im Wirbel mit sich fortführen? Nun wurde auch ich in **Geschäfte** verwickelt, die mir das um fo viel wahrcheinlicher machten. [...]

Kein Highlighting

Geschäfte

Suchen

Laden.....

Seite 1

ihn im Wirbel mit sich fortführen? Nun wurde auch ich in **Gefchäfte** verwickelt, die mir das um fo viel wahrcheinlicher machten.

schäften mit sich.
(Fortsetzung, f. Nr. 30)

Saarfelds Posten erforderte öftere Reisen in den Provinzen; er war mit Geschäften überhäuft; er hielt um wichtige Stellen an; eine glänzende Laufbahn ihm; ich merkte bald, daß seine Seele anfing, von Ehrgeiz ergriffen zu werden; seine Ausdrücke verloren nach und nach etwas von ihrer Kraft; seine Briefe wurden immer kürzer und kürzer; es fielen schon Posttage aus; er sprach nur noch bloß vom Fräulein von B., wenn die Coquette erwähnte, und dies geschah gar nicht im gewohnten leidenschaftlichen Ton. Mir wurde immer deutlicher; ich las in seiner Seele; eine Leidenschaft verdrängte das erste. Seine Reize hatten eine andre Richtung genommen, und ich weiß nicht, war es Abneigung oder Stolz; genug, ich ahmte Saarfeldern nach, schrieb immer seltener und kürzer; und um die letzte Herabsetzung zu vermeiden, schrieb ich endlich gar nicht mehr an den falschen Höfling. Er beklagte sich auch keinesweges darüber. Es gieng mir herzlich nahe, aber ich unterbrach das Stillschweigen nicht. In meinem Herzen entschuldigte ich ihn: vielleicht hat er heimlichen Kummer, sagte ich; wer weiß, wie viele Geschäfte

Highlighting

Geschäfte

Suchen

Laden.....

Seite 1

ihn im Wirbel mit sich fortführen? Nun wurde auch ich in **Geschäfte** verwickelt, die mir das um fo viel wahrscheinlicher machten.

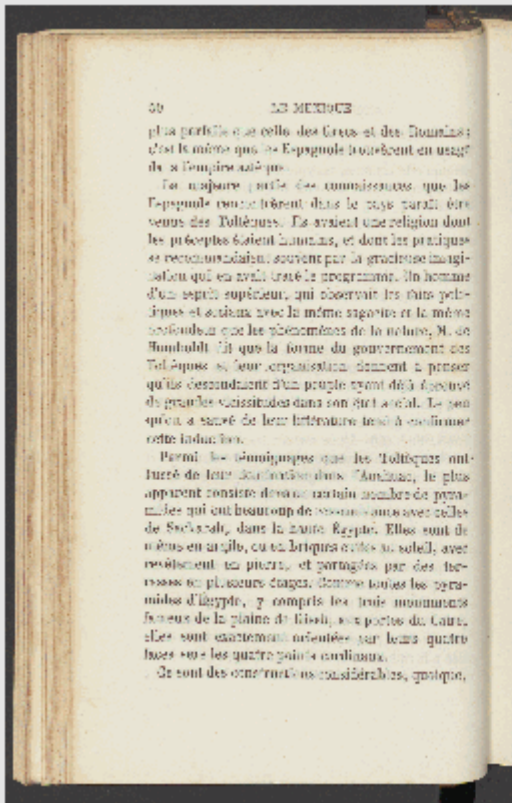
Geschäften mit sich.
(Fortsetzung, f. Nr. 30)

Saarfelds Posten erforderte öftere Reisen in den Pro:
war mit **Geschäften** überhäuft; er hielt um
wichtige Stellen an; eine glänzende Laufbahn
hm; ich merkte bald, daß seine Seele anfing,
von Ehrgeiz ergriffen zu werden; seine Ausdrücke verlor
ren nach und nach etwas von ihrer Kraft; seine Briefe
wurden immer kürzer und kürzer; es fielen schon Posttage
aus; er sprach nur noch bloß vom Fräulein von B., wenn
eroinette erwähnte, und dies geschah gar nicht
im gewohnten leidenschaftlichen Ton. Mir
immer deutlicher; ich las in seiner Seele; eine
Leidenschaft verdrängte das erste. Seine Reiz
gaben hatte eine andre Richtung genommen, und ich
weiß nicht, war es Abneigung oder Stolz; genug, ich ahm
te Saarfeldern nach, schrieb immer seltener und kürzer;
und um die letzte Herabsetzung zu vermeiden, schrieb ich
endlich gar nicht mehr an den falschen Höfling. Er beklagte
sich auch keinesweges darüber. Es gieng mir herzlich na
he, aber ich unterbrach das Stillschweigen nicht. In weis
nem Herzen entschuldigte ich ihn: vielleicht hat er heime
lichen Kummer, sagte ich; wer weiß, wie viele **Geschäfte**

Experiment: ICU Transformation

```
<fieldType name="text_ocr" class="solr.TextField" storeOffsetsWithPositions="true" termVectors="true">
  <analyzer type="index">
    <!-- For converting OCR to plaintext -->
    <charFilter class="solr.OCRCharFilterFactory" />
    <!-- michaelkubina: account for some ocr-engines escaping html characters -->
    <charFilter class="solr.HTMLStripCharFilterFactory"/>
    <!-- michaelkubina: use character filter for special character mapping -->
    <charFilter name="mapping" mapping="characters.txt"/>
    <!-- michaelkubina: tokenize at whitespace -->
    <tokenizer class="solr.WhitespaceTokenizerFactory"/>
    <!-- michaelkubina: transliterate according to ICU Transformation to Latin script -->
    <filter class="solr.ICUTransformFilterFactory" id="Any-Latin"/>
    <!-- michaelkubina: apply ICU Folding on Latin script (basically like ascii folding) -->
    <filter name="icuFolding"/>
    <!-- michaelkubina: Lowercase tokens as soon as possible -->
    <filter class="solr.LowerCaseFilterFactory"/>
    <!-- michaelkubina: compound tokens if hyphen at the end of one token suggests it being part of a compound word with the then following token -->
    <filter class="solr.HyphenatedWordsFilterFactory"/>
    <!-- michaelkubina: catenate hyphenated words or combinations of alphanumericals ; camelcase wont happen due to Lowercasefilter at the beginning;
    <filter class="solr.WordDelimiterGraphFilterFactory" generateWordParts="1" preserveOriginal="0" generateNumberParts="1" catenateWords="1" catenate
    <!-- michaelkubina: flatten word graph -->
    <filter class="solr.FlattenGraphFilterFactory"/>
    <!-- michaelkubina: keep keywords as duplicate tokens and prevent them from getting stemmed -->
    <filter class="solr.KeywordRepeatFilterFactory"/>
    <!-- michaelkubina: remove any trailing or leading whitespaces from tokens, if it happened for any reason -->
    <filter class="solr.TrimFilterFactory"/>
    <!-- michaelkubina: do the stemming -->
    <filter class="solr.SnowballPorterFilterFactory" language="German" protected="protwords.txt"/>
    <!-- michaelkubina: reverse all tokens, so that they can be found faster in a reverse wildcard search (only needed at index-time) -->
    <filter class="solr.ReversedWildcardFilterFactory" withOriginal="true" maxPosAsterisk="3" maxPosQuestion="2" maxFractionAsterisk="0.33"/>
    <!-- michaelkubina: remove duplicate tokens for the same position increment -->
```

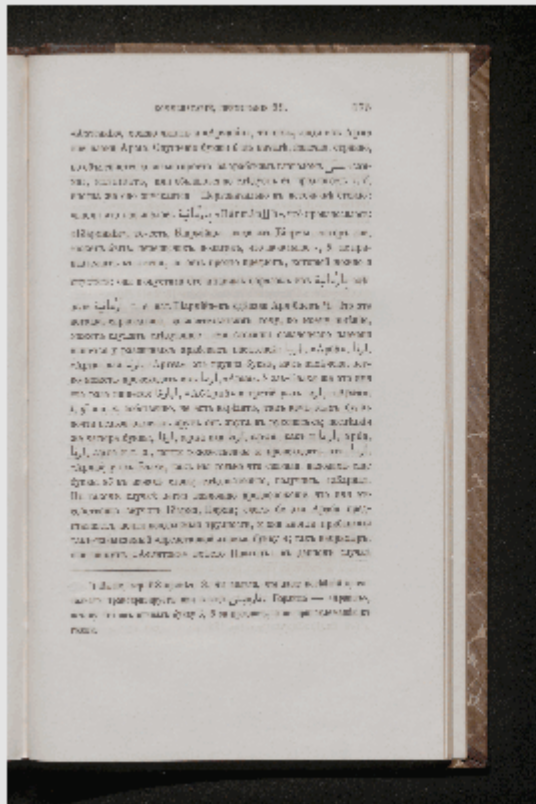
Query: Platon



Seite 48
Titel Le Mexique ancien et moderne

[...] pagnol , Ixtlixochitl, sont d'une rare beauté. Quant à ses idées religieuses, c'est à croire qu'il avait con versé avec **Platon** ou avec saint Paul. Après avoir regagné le trône de ses pères, il accorda une amnis tie générale en prononçant ces paroles : « Un roi [...]

auch Query: Platon



Seite 193
Titel Izvestija o Hozarach, Burtasach, Bolgarach, Madjarach Slavjanach i Russach

[...] ставляетъ почти неодолимья трудности, и они иногда прибавляли такъ-называемый «представной элифъ», букву а; такъ напримѣръ, они пишутъ «Афлатонъ» вмѣсто **Платонъ**; въ данномъ случаѣ °) Выше, стр. 58 прпмѣч. 8, мы видѣли, что даже новѣйшій ориен талисть транскрипируетъ имя города Барайшь — «Арайшь», [...]

Platon
Πλάτων
Платон

aber nicht:

ⲠⲓⲮⲉⲭ
أفلاطون

Fazit


- Hoher Stellenwert in Forschung und Lehre
- Darstellung in Kitodo.Presentation prinzipiell möglich
- Schemata berücksichtigen Sprache und Schrift
- Umdenken erforderlich
 - Erschließungspraxis in diesem Fall oft unzureichend
 - Gemischtsprachliche/-schriftliche Texte nicht als Sonderfall sondern als Regelfall betrachten
- Akut: Zeichengenaue OCR deutschsprachiger Texte
 - potentielle Treffer mit æ, œ, ů und ı gehen im Suchraum verloren
 - Highlighting zerbrechlich in diesem Fall
- Helper::getScriptName parallel zur ISO-639 auch für ISO-15924

Vielen Dank

Michael Kubina

Von-Melle-Park 3
20146 Hamburg

040 / 4 28 38-5586
michael.kubina@sub.uni-hamburg.de

 www.sub.uni-hamburg.de/
 www.facebook.com/stabihh
 openbiblio.social/@stabihh