STAATS- UND UNIVERSITÄTS-
BIBLIOTHEK
HAMBURG
CARL VON OSSIETZKY

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

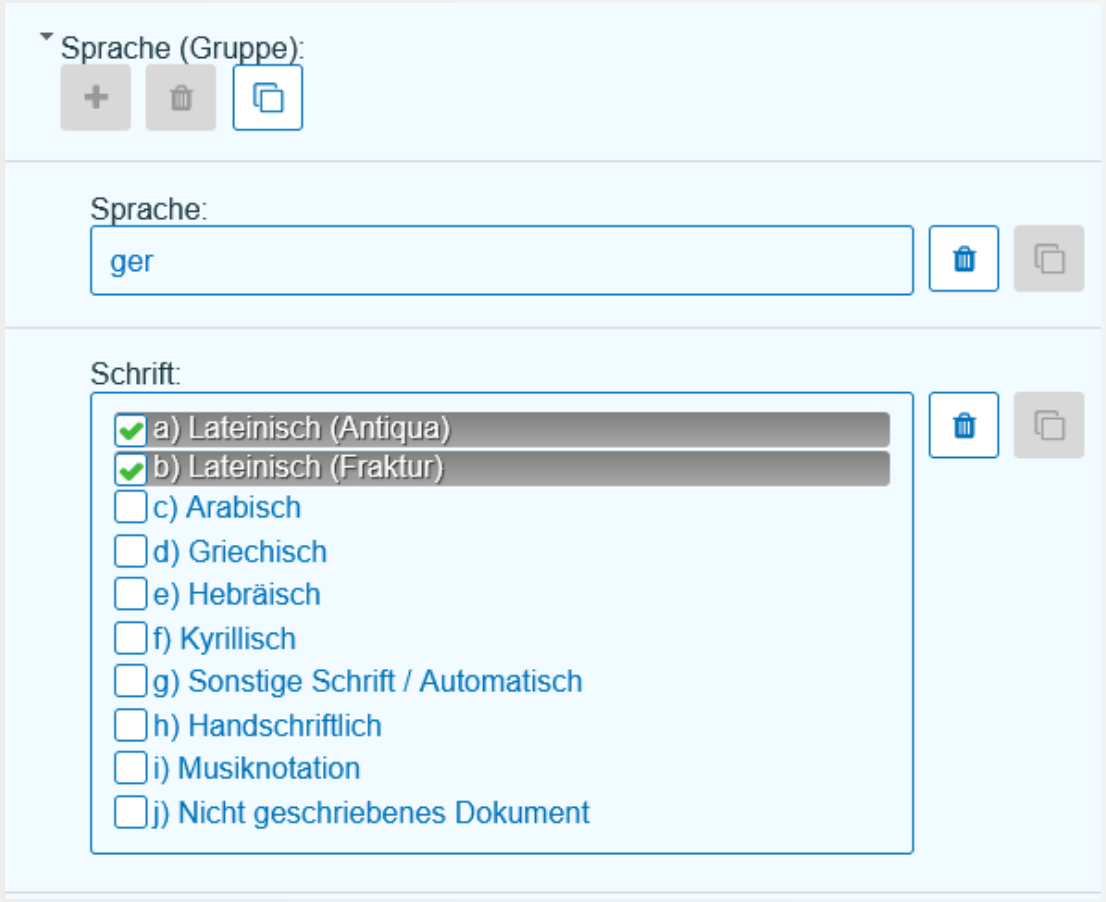
```
<selectl tns:ref="DocLanguage">
    <item tns:selected="false">
        <label>Arabisch</label>
        <value>ara</value>
    </item>
    <item tns:selected="false">
        <label>Aram&#228;isch</label>
        <value>arc</value>
    </item>
    <item tns:selected="false">
        <label>Armenisch</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>
```

```
<selectl 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>
</selectl>
```

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



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-Andwenungsprofil(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/dfgviewer/MODS-Anwendungsprofil_2.3.1.pdf

Use Case: Automatic Text Recognition

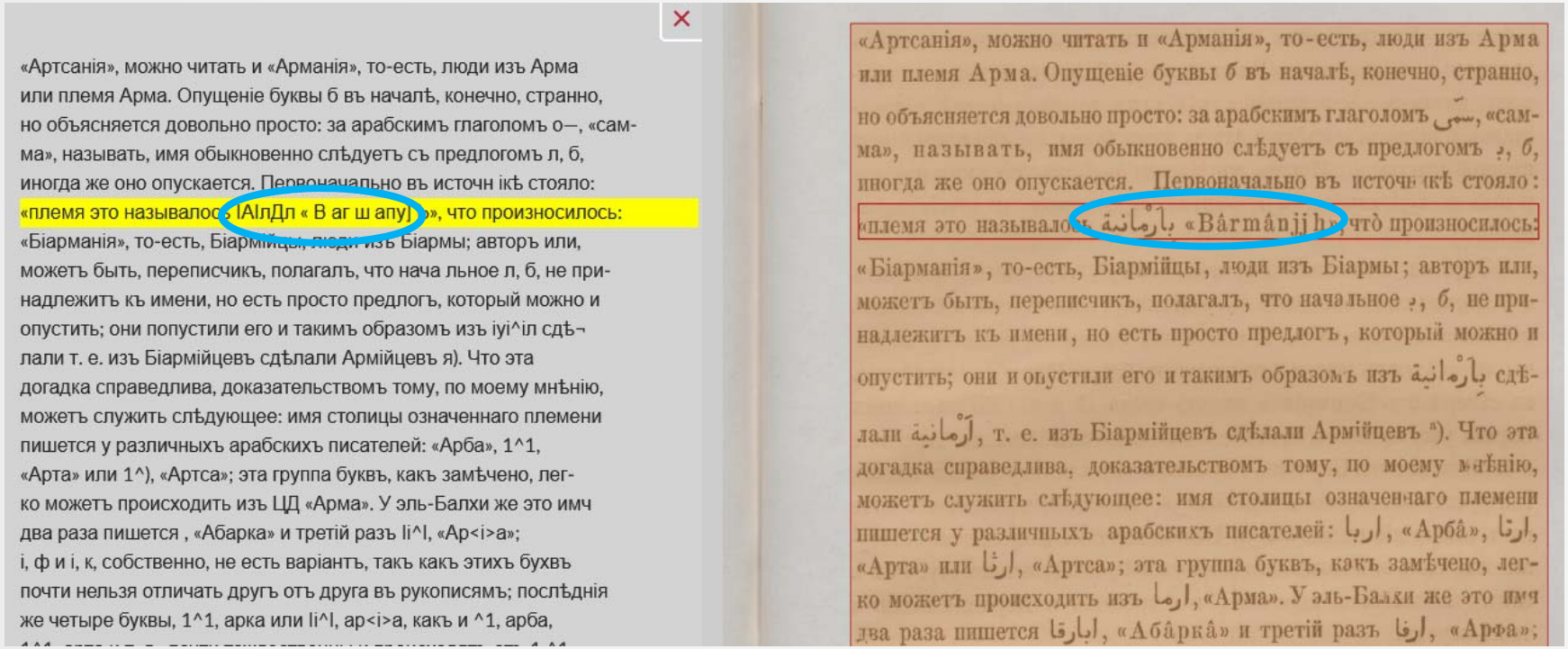
```
# wahl des schriftmodels durch die in kitodo ausgewaehlte schrift
case $2 in
    "Latn")
        TESSMODEL+="+script/Latin"
        ;;
    "Latf")
        TESSMODEL+="+script/Fraktur"
        ;;
    "Arab")
        TESSMODEL+="+script/Arabic"
        ;;
    "Grek")
        # spezialmodel (alt-)griechisch
        TESSMODELPRIO+="+external/grc_hist"
        # fallback auf griechisch
        TESSMODEL+="+script/Greek"
        ;;
    "Hebr")
        TESSMODEL+="+script/Hebrew"
        ;;
    "Cyrl")
        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 Cyrl Latn" territories="AZ"/>
<language type="az" territories="IQ IR RU" alt="secondary"/>
<language type="ba" scripts="Cyrl"/>
<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 Cyrl Latn"
```

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

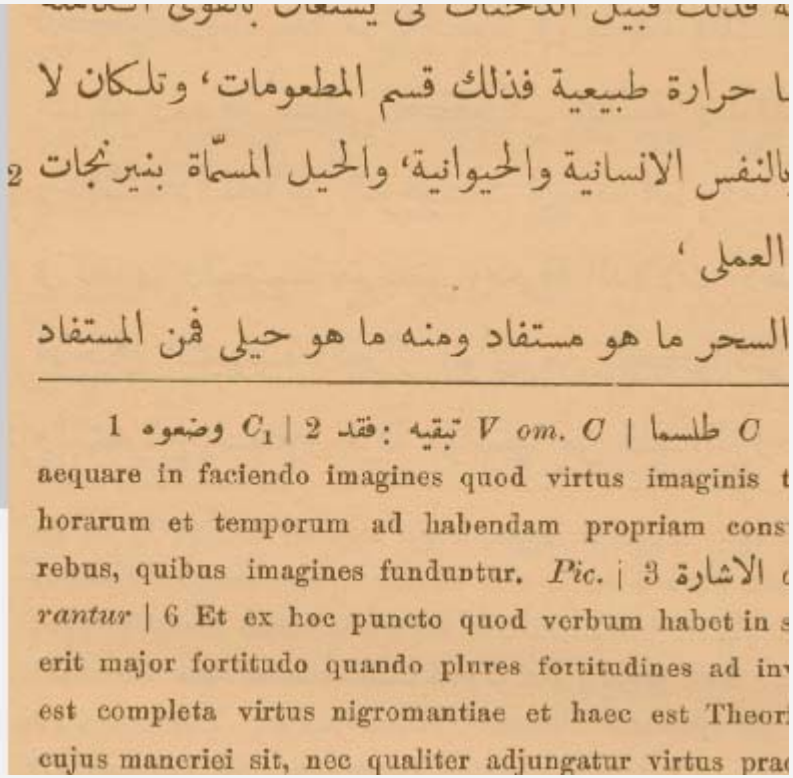
Problem: Sprach- & Schriftmix



«Артсанія», можно читать и «Арманія», то-есть, люди изъ Арма
или племя Арма. Опущеніе буквы б въ началѣ, конечно, странно,
но объясняется довольно просто: за арабскимъ глаголомъ о—, «сам-
ма», называть, имя обыкновенно слѣдуетъ съ предлогомъ л, б,
иногда же оно опускается. Первоначально въ источн ікѣ стояло:
«племя это называлось IAIлДл « В аг ш апу ». что произносилось:
«Біарманія», то-есть, Біармійцы, люди изъ Біармы; авторъ или,
можетъ быть, переписчикъ, полагалъ, что нача льное л, б, не при-
надлежитъ къ имени, но есть просто предлогъ, который можно и
опустить; они попустили его и такимъ образомъ изъ іуі^іл сдѣ¬
лали 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», 1^1,
«Арта» или 1^), «Артса»; эта группа буквъ, какъ замѣчено, лег-
ко можетъ происходить изъ ЦД «Арма». У эль-Балхи же это имч
два раза пишется , «Абарка» и третій разъ Ii^I, «Ар<i>а»;
і, ф и і, к, собственно, не есть варіантъ, такъ какъ этихъ бухвъ
почти нельзя отличать другъ отъ друга въ рукописямъ; послѣднія
же четыре буквы, 1^1, арка или Ii^I, ар<i>а, какъ и ^1, арба,

«Артсанія», можно читать и «Арманія», то-есть, люди изъ Арма
или племя Арма. Опущеніе буквы *б* въ началѣ, конечно, странно,
но объясняется довольно просто: за арабскимъ глаголомъ سمّى, «сам-
ма», называть, имя обыкновенно слѣдуетъ съ предлогомъ ب, *б*,
иногда же оно опускается. Первоначально въ источн ікѣ стояло:
«племя это называлось بارْمانية «Bârmânjjh», что произносилось:
«Біарманія», то-есть, Біармійцы, люди изъ Біармы; авторъ или,
можетъ быть, переписчикъ, полагалъ, что начальное ب, *б*, не при-
надлежитъ къ имени, но есть просто предлогъ, который можно и
опустить; они и опустили его и такимъ образомъ изъ بارْمانية сдѣ-
лали اَرْمانية, т. е. изъ Біармійцевъ сдѣлали Армійцевъ [a]). Что эта
догадка справедлива, доказательствомъ тому, по моему мнѣнію,
можетъ служить слѣдующее: имя столицы означеннаго племени
пишется у различныхъ арабскихъ писателей: اربا, «Арба̂», ارتا,
«Арта» или ارثا, «Артса»; эта группа буквъ, какъ замѣчено, лег-
ко можетъ происходить изъ ارما, «Арма». У эль-Балхи же это имя
два раза пишется ابارقا, «Абâркâ» и третій разъ ارفا, «Арфа»;

Problem: Sprach- & Schriftmix



Problem: Zeichengenaue OCR

å ů

ȯ ɾ

¯_(ツ)_/¯

Aktuelle schema.xml

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

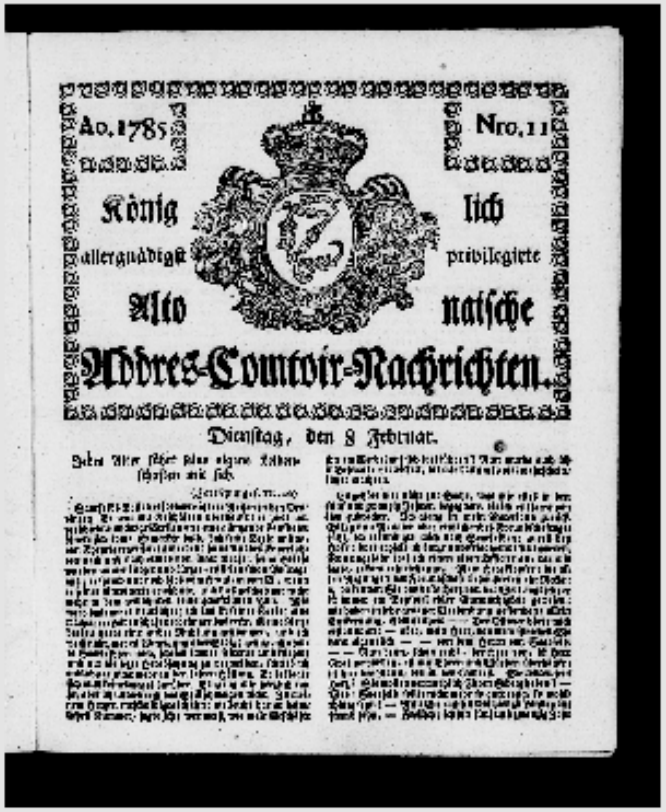
Lösung? Character Folding?

Analyzer	Abkürzung	Index	Query
OCR Character Filter	OCF	Geſchåfte	
HTML Special Character Filter	HTMLSCF	Geſchåfte	
Whitespace Tokenizer	WT	Geſchåfte	Geschäft
AsciiFoldingFilter	ASCIIFF	Geschåfte	Geschaft
SynonymGraphFilter	SGF	Geschåfte	Geschaft
FlattengraphFilter	FGF	Geschåfte	
Stop(word) Filter	SF	Geschåfte	Geschaft
WordDelimiterGraphFilter	WDGF	Geschåfte	Geschaft
ReverseWildcardSearchFilter	RWF	Geschåfte	
LowercaseFilter	LCF	geschåfte	geschaft
TrimFilter	TF	geschåfte	geschaft
SnowballFilter	SF	geschåft	geschaft
RemoveDuplicateTokenFilter	RDTF	geschåft	geschaft

Lösung? Character Mapping?

Analyzer	Abkürzung	Index	Query
OCR Character Filter	OCF	Geſchåfte	
HTML Special Character Filter	HTMLSCF	Geſchåfte	
MappingCharacterFilter	MCF	Geſchäfte	Geschäft
Whitespace Tokenizer	WT	Geſchäfte	Geschäft
AsciiFoldingFilter	ASCIIFF	Geschafte	Geschaft
SynonymGraphFilter	SGF	Geschafte	Geschaft
FlattengraphFilter	FGF	Geschafte	
Stop(word) Filter	SF	Geschafte	Geschaft
WordDelimiterGraphFilter	WDGF	Geschafte	Geschaft
ReverseWildcardSearchFilter	RWF	Geschafte	
LowercaseFilter	LCF	geschafte	geschaft
TrimFilter	TF	geschafte	geschaft
SnowballFilter	SF	geschaft	geschaft
RemoveDuplicateTokenFilter	RDTF	geschaft	geschaft

Noch nicht ganz …


Ao. 1785 Nro. 11

Königlich allergnädigst privilegirte

Altonaische

Addreß-Comtoir-Nachrichten.

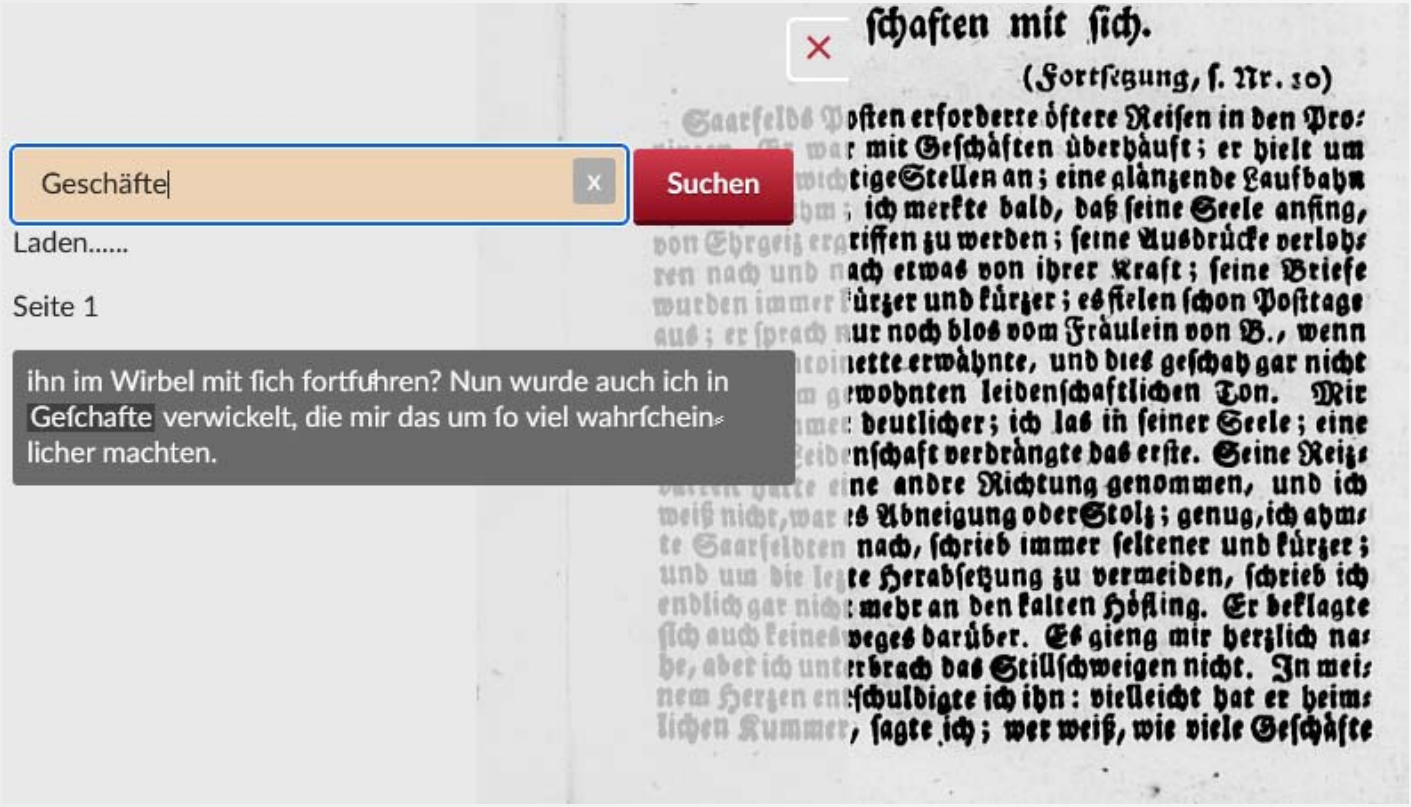
Dienstag, den 8 Februar.

Seite 1

Titel Königlich privilegirte Altonaer Adreß-Comtoir-Nachrichten

[...] ihn im Wirbel mit ſich fortführen? Nun wurde auch ich in *Geſchafte* verwickelt, die mir das um ſo viel wahrſchein- licher machten. [...]

Kein Highlighting

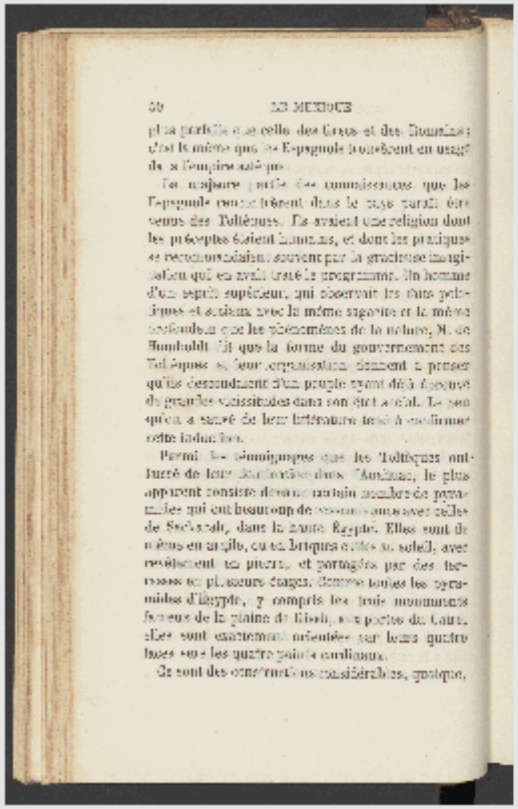


Highlighting

Experiment: ICU Transformation

```
<fieldType name="text_ocr" class="solr.TextField" storeOffsetsWithPositions="true" termVectors="true">
        <analyzer type="index">
                <!-- For converting OCR to plaintext -->
                <charFilter class="solrocr.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: translitarate 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

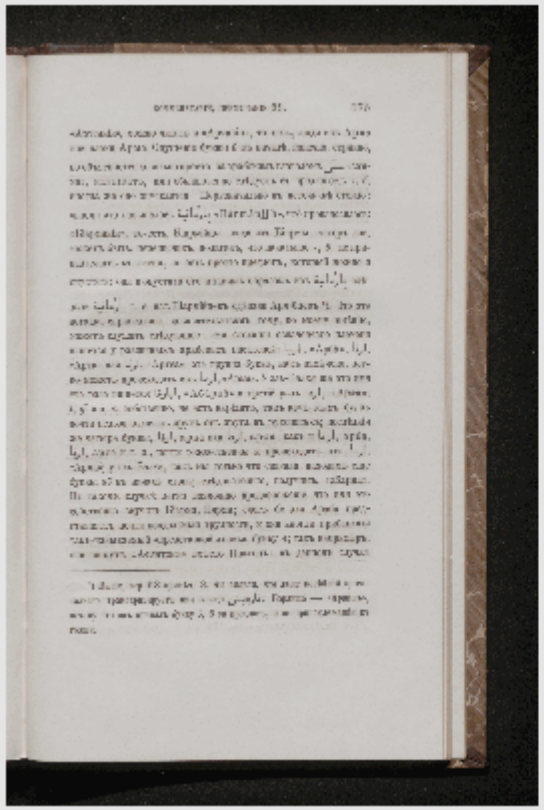


Seite 48

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 неодолимыя 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 оріен талистъ транскрипируетъ имя города Бараишъ — «Араишъ», [...]

Platon

Πλάτων

Платон

aber nicht:

אפלטון

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, uᵉ und ſ 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