

AN OPEN-SOURCE FRAMEWORK FOR INTEGRATING MULTI-SOURCE LAYOUT AND TEXT

RECOGNITION TOOLS INTO SCALABLE OCR WORKFLOWS

KAY-MICHAEL WÜRZNER

KONSTANTIN BAIERER

OVERVIEW

1. The situation
2. OCR-D process
3. Open source framework
4. Multi-source tools
5. Scalable workflows

USERS WANT TEXT DATA



MASSIVE AMOUNTS


```
</Shape>
<TextLine ID="tl_5" BASELINE="860" HEIGHT="55" WIDTH="564" VPOS="805" HPOS="233" STYLEREFS="Arial_12.0_-----">
  <String ID="w_w1aab1b3b2b3b1ab1" HEIGHT="51" WIDTH="306" VPOS="807" HPOS="233" STYLEREFS="Arial_12.0_-----" CONTENT="Beantwortung"/>
  <String ID="w_w1aab1b3b2b3b1ac27" HEIGHT="38" WIDTH="61" VPOS="810" HPOS="569" STYLEREFS="Arial_12.0_-----" CONTENT="der"/>
  <String ID="word_1478541289590_808" HEIGHT="53" WIDTH="124" VPOS="806" HPOS="654" STYLEREFS="Arial_12.0_-----" CONTENT="Frage"/>
  <String ID="word_1478541289588_807" HEIGHT="53" WIDTH="19" VPOS="806" HPOS="778" STYLEREFS="Arial_12.0_-----" CONTENT=":"/>
</TextLine>
<TextLine ID="tl_6" BASELINE="941" HEIGHT="55" WIDTH="683" VPOS="886" HPOS="177" STYLEREFS="-----">
  <String ID="w_w1aab1b3b2b3b3ab1" HEIGHT="46" WIDTH="139" VPOS="888" HPOS="177" STYLEREFS="Arial_12.0_-----" CONTENT="Was"/>
  <String ID="w_w1aab1b3b2b3b3ab9" HEIGHT="51" WIDTH="56" VPOS="890" HPOS="362" STYLEREFS="Arial_12.0_-----" CONTENT="ift"/>
  <String ID="word_1478541293583_810" HEIGHT="52" WIDTH="367" VPOS="887" HPOS="465" STYLEREFS="Arial_12.0_-----" CONTENT="Aufklärung"/>
  <String ID="word_1478541293581_809" HEIGHT="52" WIDTH="28" VPOS="887" HPOS="832" STYLEREFS="Arial_12.0_-----" CONTENT="?"/>
</TextLine>
</TextBlock>
<TextBlock ID="r_2_3" HEIGHT="42" WIDTH="474" VPOS="977" HPOS="277">
  <Shape>
    <Polygon POINTS="277,977 751,977 751,1019 277,1019"/>
  </Shape>
</TextBlock>
a/kant_aufklaerung_1784/alto/kant_aufklaerung_1784_0017.xml xml ≡ 22%
</mets:mdWrap>
<mets:dmdSec>
<mets:fileSec>
<mets:fileGrp USE="OCR-D-IMG">
  <mets:file ID="INPUT_0020" MIMETYPE="image/tiff">
    <mets:Flocat LOCTYPE="URL" xlink:href="https://github.com/OCR-D/assets/raw/master/data/kant_aufklaerung_1784/kant_aufklaerung_1784_0017.tif" />
  </mets:file>
</mets:fileGrp>
<mets:fileGrp USE="OCR-D-GT-PAGE">
  <mets:file ID="PAGE_0020_PAGE" MIMETYPE="application/vnd.prima.page+xml">
    <mets:Flocat LOCTYPE="URL" xlink:href="https://github.com/OCR-D/assets/raw/master/data/kant_aufklaerung_1784/page/kant_aufklaerung_1784_0020.xml" />
  </mets:file>
</mets:fileGrp>
<mets:fileGrp USE="OCR-D-GT-ALTO">
  <mets:file ID="ALTO_0020_ALTO" MIMETYPE="application/vnd.prima.alto+xml">
    <mets:Flocat LOCTYPE="URL" xlink:href="https://github.com/OCR-D/assets/raw/master/data/kant_aufklaerung_1784/alto/kant_aufklaerung_1784_0020.xml" />
  </mets:file>
</mets:fileGrp>
</mets:fileSec>
</mets:dmdSec>
</mets:mdWrap>
</mets:package>
assets/data/kant_aufklaerung_1784/mets.xml
```

STRUCTURED

Filter

HistoryCollections

Clear all

▼ Today

GET

http://api.bart.gov/api/bsa.aspx?cmd=bsa&key=MW9S-E7SL-26DU-VV8V&date=today

POST

https://api.imgur.com/oauth2/token

GET

http://api.giphy.com/v1/gifs/search?q={{emotion}}+{{animal}}&api_key=dc6zaTOxFjnzC

GET

https://api.forecast.io/forecast/{{forecastapikey}}/{{latlong}}

GET

https://maps.googleapis.com/maps/api/geocode/json?address="1 Infinite Loop, Cupertino, CA 95014"&key=AIzaSyB63R12ZTUfBUQMUMZKUDFwdMPn66mQrg4p

POST

https://hooks.slack.com/services/T02G7V5JE/B63R12ZTUfBUQMUMZKUDFwdMPn66mQrg4p

GET

http://api.bart.gov/api/etd.aspx?cmd=etd&orig=POWL&key=MW9S-E7SL-26DU-VV8V&dir=n

GET

http://api.bart.gov/api/etd.aspx?cmd=etd&orig=MONT&key=MW9S-E7SL-26DU-VV8V&dir=n

GET

http://api.bart.gov/api/sched.aspx?cmd=depart&orig=POWL&dest=ORIN&b=0&a=2&l=1&key=MW9S-E7SL-26DU-VV8V&date=today

GET

http://api.bart.gov/api/sched.aspx?cmd=depart&orig=MONT&dest=ORIN&b=0&a=2&l=1&key=MW9S-E7SL-26DU-VV8V&date=today

GET

http://api.bart.gov/api/bsa.aspx?cmd=bsa&key=MW9S-E7SL-26DU-VV8V&date=today

GET

https://www.random.org/cgi-bin/ra

https://www.random.org/cgi-bin/ra

Bart Advisories

+...

▶ Bart Advisories

Examples (0) ▼

GET

http://api.bart.gov/api/bsa.aspx?cmd=bsa&key=MW9S-E7SL-26DU-VV8V&date=today

Params

Send

Save

Authorization

Headers

Body

Pre-request Script

Tests

TYPE

Inherit auth from parent

This request is using an authorization helper from collection [Kasey's Bart Checker](#).

Body

Cookies

Headers (12)

Test Results (1/1)

Status: 200 OK

Time: 166 ms

Size: 674 B

Pretty

Raw

Preview

XML

Save Response

1

<?xml version="1.0" encoding="UTF-8" ?>

2

<root>

3

<uri>

4

<![CDATA[http://api.bart.gov/api/bsa.aspx?cmd=bsa&date=today]]>

5

</uri>

6

<date>02/08/2018</date>

7

<time>14:18:00 PM PST</time>

8

<bsa>

9

<station></station>

10

<description>

11

<![CDATA[No delays reported.]]>

12

</description>

13

<sms_text>

14

<![CDATA[No delays reported.]]>

15

</sms_text>

16

</bsa>

17

<message></message>

18

</root>

EASILY ACCESSIBLE

3.4

LIBRARIES PROVIDE TEXT DATA

92749 Suchergebnisse mit folgenden Kriterien:

Kategorien

Historische Drucke ✕

erfasser in 48 Kategorien

Historische Drucke (92749)

D18 digital (26341)

heologie (22615)

Rechtswissenschaft (17651)

prachen / Literaturen (13902)

eschichte / Ethnographie / Geographie

260)

rieg 1914-1918 (7587)

stasiatica (7302)

llgemeines / Wissenschaftskunde /

teraturgeschichte (5580)

usik (4997)

inblattdrucke (3094)



Gesetzsammlung für die Königlichen Preussischen Staaten

1865. 1865 | Zeitschriftenband

Gr 3602-1865 | In: **Gesetzsammlung für die Königlichen Preussischen Staaten**



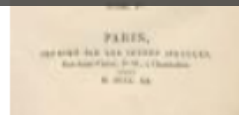
Guillié, Sébastien

LARGE AMOUNTS

Bibliothèque Ophtalmologique, Ou Recueil D'Observations Sur Les Maladies Des Yeux,

T. 1. 1820 | Band

Km 305-1,1/4 | In: **Bibliothèque Ophtalmologique, Ou Recueil D'Observations Sur Les Maladies Des Yeux, Faites A La Clinique De L'Institution Royale Des Jeunes Aveugles**



Der Herold

1839. 1839 | Zeitschriftenband

Gq 14350;Beil.3-1839 | In: **Der Herold**





[5] -- ▾



3

4

5

6

7



Gesetz-Sammlung für die Königlichen Preussischen Staaten. 1865.

Enthält die Gesetze, Verordnungen re. vom 2. Januar bis zum 28. Dezember 1865.,
nebst einigen Verordnungen re. aus den Jahren 1862. und 1864.

(Von Nr. 5991. bis Nr. 6235.)

Nr. 1. bis incl. 57.

Berlin,

zu haben im vereinigten Gesetz-Sammlungs-Debits- und Zeitungs-Komtoir.

UNSTRUCTURED

ZBIÓR PRAW DLA PAŃSTW KRÓLESTWA PRUSKIEGO. 1865.

Zawiera ustawy, rozporządzenia i t. d. od 2. Stycznia aż do 28. Grudnia 1865.,
wraz z niektórymi rozporządzeniami i t. d. z roku 1862. i 1864.

(Od No. 5991. aż do No. 6235.)

No. 1. aż włącznie do 57.

W BERLINIE,

do nabycia w połączonej biurze sprzedaży Zbioru praw i gazet.

Gesetz-Sammlung

für die , ,

Königlichen Preussischen Staaten.

1865.

Enthält die Gesetze, Verordnungen re. vom 2. Januar bis ^ 28-Dezemb 1865.,

„ebst einigen Verordnungen k. aus den Jahren 1862, und «**

(Bon Nr. 5991. bis Nr. 6235.)

Nr. 1. bis incl. 57.

zu

haben im vereinigten Gesetz-Sammlungs-Debits- und Zeitungs-Komtoir.

ZBIÓR PRAW

DLA

PANSTW

```
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002083.png"
le><mets:file ID="FILE_2084_THUMBS" MIMETYPE="image/png"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="h
.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002084.png"/></mets:file><mets:file ID="FILE_2085_THUMBS" MIMETYPE="image/png"><mets:FL
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002085.png"
le><mets:file ID="FILE_2086_THUMBS" MIMETYPE="image/png"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="h
.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002086.png"/></mets:file><mets:file ID="FILE_2087_THUMBS" MIMETYPE="image/png"><mets:FL
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002087.png"
le><mets:file ID="FILE_2088_THUMBS" MIMETYPE="image/png"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="h
.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002088.png"/></mets:file><mets:file ID="FILE_2089_THUMBS" MIMETYPE="image/png"><mets:FL
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002089.png"
le><mets:file ID="FILE_2090_THUMBS" MIMETYPE="image/png"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="h
.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002090.png"/></mets:file><mets:file ID="FILE_2091_THUMBS" MIMETYPE="image/png"><mets:FL
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002091.png"
le><mets:file ID="FILE_2092_THUMBS" MIMETYPE="image/png"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="h
.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002092.png"/></mets:file><mets:file ID="FILE_2093_THUMBS" MIMETYPE="image/png"><mets:FL
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002093.png"
le><mets:file ID="FILE_2094_THUMBS" MIMETYPE="image/png"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="h
.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002094.png"/></mets:file><mets:file ID="FILE_2095_THUMBS" MIMETYPE="image/png"><mets:FL
link="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://content.staatsbibliothek-berlin.de/dms/PPN746523211/150/0/00002095.png"
le></mets:fileGrp><mets:fileGrp USE="FULLTEXT"><mets:file ID="FILE_0001_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat xmlns:xlink="http://www.w
link="http://digital.staatsbibliothek-berlin.de/ocrresolver/PPN746523211/00000001.xml"/></mets:file><mets:file
ID="FILE_0002_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://digital.
bibliothek-berlin.de/ocrresolver/PPN746523211/00000002.xml"/></mets:file><mets:file ID="FILE_0003_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat
http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://digital.staatsbibliothek-berlin.de/ocrresolver/PPN746523211/00000003.xml"/></
mets:file ID="FILE_0004_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http:
.staatsbibliothek-berlin.de/ocrresolver/PPN746523211/00000004.xml"/></mets:file><mets:file ID="FILE_0005_FULLTEXT" MIMETYPE="text/xml"><mets
xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://digital.staatsbibliothek-berlin.de/ocrresolver/PPN746523211/0000
/mets:file><mets:file ID="FILE_0006_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlin
http://digital.staatsbibliothek-berlin.de/ocrresolver/PPN746523211/00000006.xml"/></mets:file><mets:file ID="FILE_0007_FULLTEXT" MIMETYPE="te
mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://digital.staatsbibliothek-berlin.de/ocrresolver/PPN746
7.xml"/></mets:file><mets:file ID="FILE_0008_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="
ref="http://digital.staatsbibliothek-berlin.de/ocrresolver/PPN746523211/00000008.xml"/></mets:file><mets:file ID="FILE_0009_FULLTEXT"
E="text/xml"><mets:Flocat xmlns:xlink="http://www.w3.org/1999/xlink" LOCTYPE="URL" xlink:href="http://digital.staatsbibliothek-berlin.de/
lver/PPN746523211/00000009.xml"/></mets:file><mets:file ID="FILE_0010_FULLTEXT" MIMETYPE="text/xml"><mets:Flocat xmlns:xlink="http://www.w3.
~/Downloads/PPN746523211.mets.xml
```

WHY THE DISCREPANCY?

- **The Reality**

- Masses of books/document are available & already scanned
- It is unclear if Antiqua and/or Gothic/Fraktur fonts are used in the documents
- Pre-Sorting is impossible, it would be too time/cost expensive

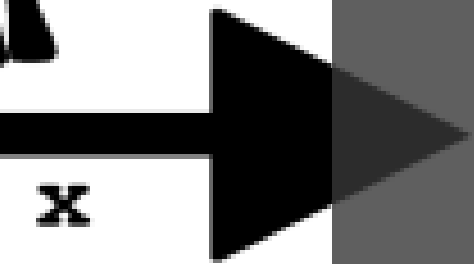
- **ABBYY Europe's Answer**

LITTLE COMPETITION

The pricing is now ready for “mass processing”

Example: Revolving credit line with non-economic

- 10.000 pages – 299 Euro – available online
- 500.000 pages* – 5.000 Euro = 1 Euro cent per page = ca 2.000 books a 250 pages
- Over 3 Mio pages* - ca 0,52 Euro cent per page = 12.000 books a 1,25 € (250 pages)
- Over 10 Mio pages* - ca. 40.000 books = ca. 0,5 € per book



ACADEMICAL SOLUTIONS NON-SUSTAINABLE

"Black Box"

INFLEXIBLE WORKFLOWS



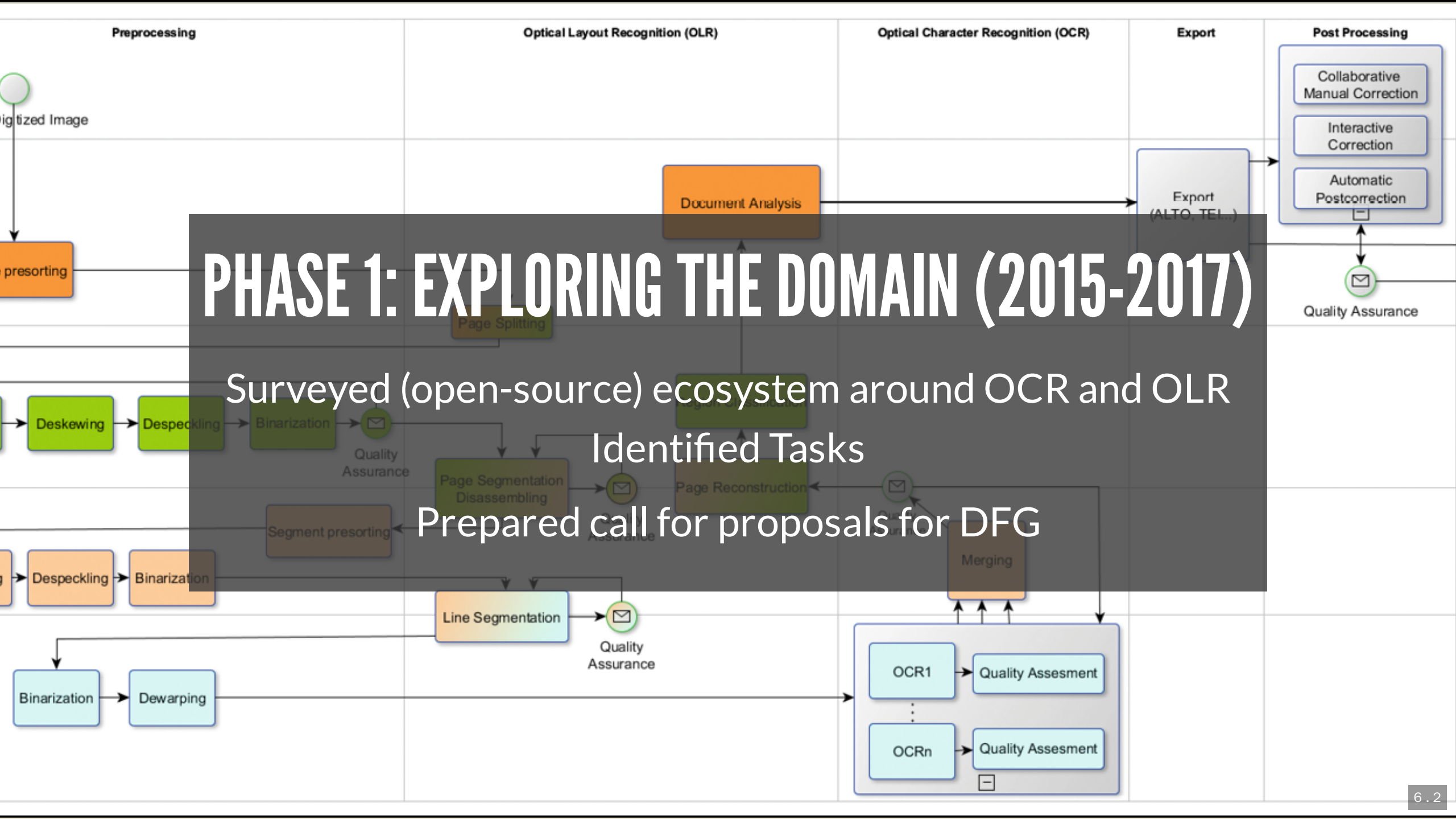
Output, 1

OCR-D COORDINATION PROJECT

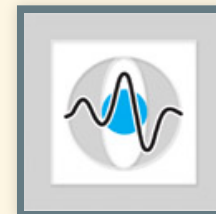


HERZOG AUGUST BIBLIOTHEK
WOLFENBÜTTEL





PHASE 2: MODULE PROJECT PROTOTYPING (2017-2018)



PHASE 3: GOING PRODUCTIVE (2018-2020)

Integrate with existing digitization workflow software, e.g.
Kitodo

Make OCR-D developed software widely deployable

Advise DFG on OCR requirements for "Praxisrichtlinien"

"MULTI SOURCE"

