Technische Universität Braunschweig

Integration von OCR-D in Kitodo: die ersten Implementierungsschritte und praktische Anwendung

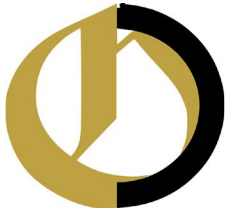
Katja Rykhlinskaya, Sven Marcus, Michael Kotzyba, Robert Strötgen

Allgemeine Projektinformationen

- **Projektförderung:**
  - DFG, Dauer 2 Jahre
- **Projektbeteiligte Institutionen:**
  - UB Braunschweig
  - SLUB Dresden
  - UB Mannheim
- **Projektaustausch:**
  - BS-intern wöchentlich
  - Projektintern monatlich
  - OCR-D-Community

KITODO

KEY TO DIGITAL OBJECTS

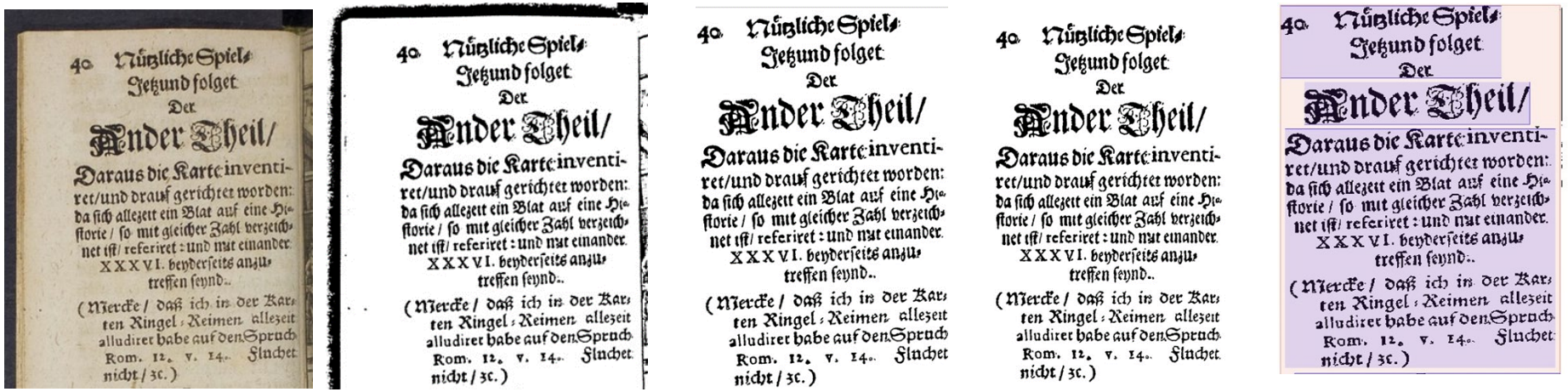


OCR-D

Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

Was macht OCR-D?

Binarisierung (aus Graustufenbild wird S/W-Bild generiert) - Zuschneiden (Ränder werden entfernt) - Entrauschen (Flecken werden entfernt) -Entzerrung (Schräglage der Seite wird korrigiert) - Entschärfung (Textzeilen werden gestreckt) - Segmentierung (Regionen werden erkannt)- Erkennung (In jedem Region wird der Text erkannt)

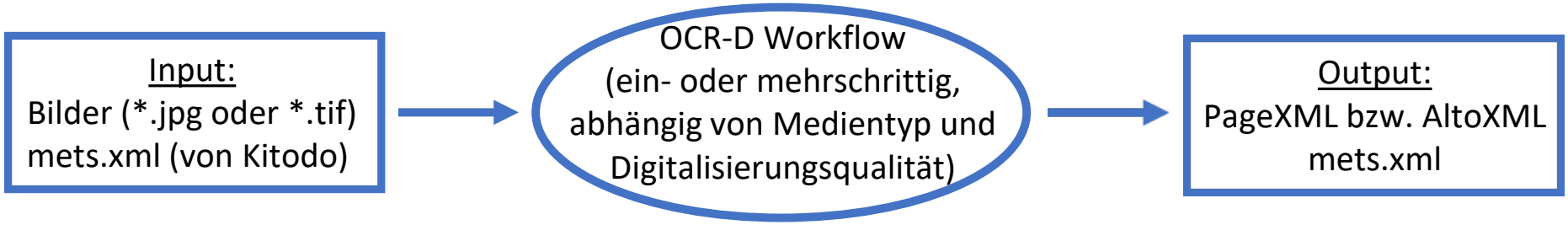


Ergebnis (Ausschnitt aus einer Alto-Datei):

```
<String ID="region0003_line0001_word0003" HEIGHT="28" WIDTH="75" HPOS="827" VPOS="370" CONTENT=„Spiel">
    <Shape>
      <Polygon POINTS="827,370 902,370 902,398 827,398"/>
    </Shape>
  </String>
```

Aktuelle OCR-D Anwendung

Die bereits digitalisierten Medien können auf dem OCR-D Server prozessiert werden!



Einschrittiger Workflow:

`ocrd-tesserocr-recognize` mit dem Model „Fraktur_GT4HistOCR“
(kann gleichzeitig Binarisierung, Regionalsegmentierung, Tabellenerkennung, Liniensegmentierung und Texterkennung durchführen).

Merhschrittiger Workflow:

`ocrd-sbb-binarize` mit dem Model „default-2021-03-09“
`ocrd-eynollah-segment` mit dem Model „default“
`ocrd-cis-ocropy-dewarp`
`ocrd-tesserocr-recognize` mit dem Model „frak2021-0.905“

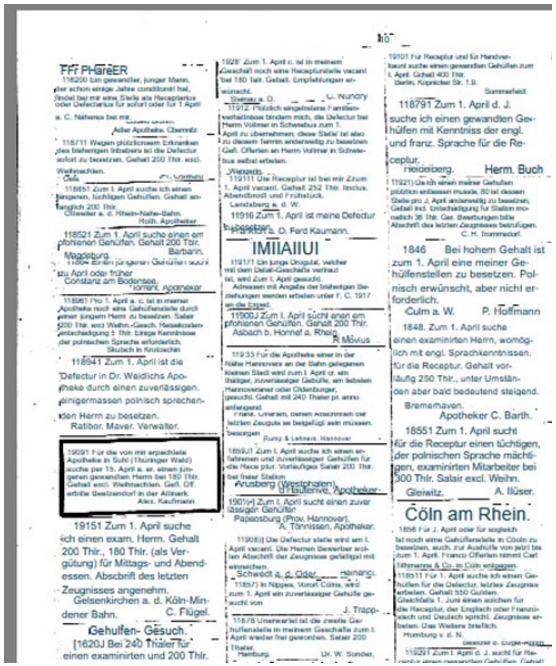
Page-to-Alto Umformung:

`ocrd-fileformat-transform`

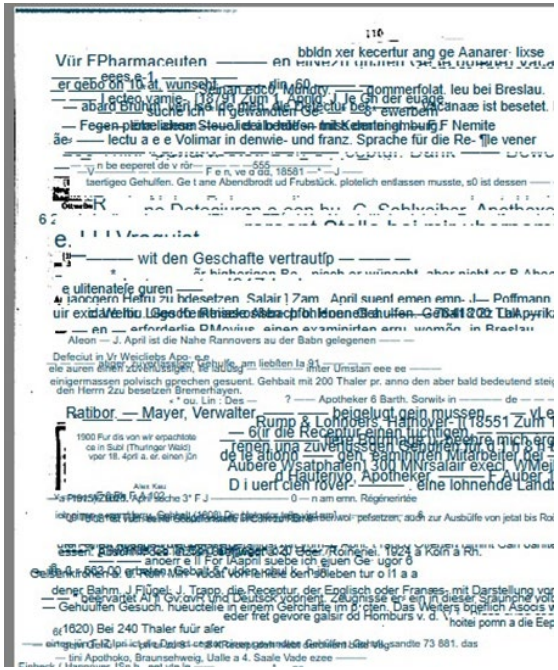
Erkennungsbeispiel 1: Pharmazeutische Zeitung J. 1872

Originalbild

Mehrschrittiger Workflow

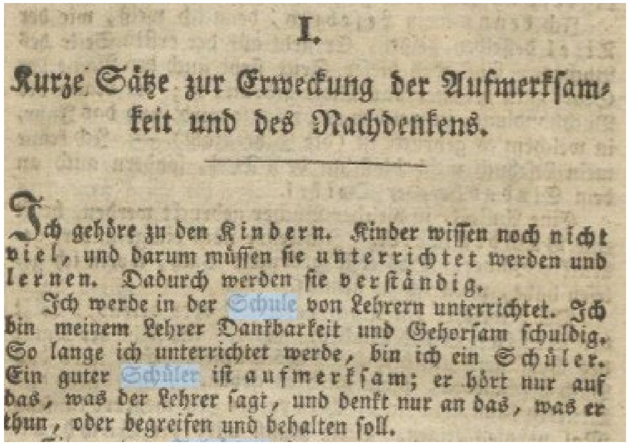


Einschrittiger Workflow



Erkennungsbeispiel 2: Kinderbuch J.1819

Originalbild


I.

Kurze Sätze zur Erweckung der Aufmerksam-
keit und des Nachdenkens.

Ich gehöre zu den Kindern. Kinder wissen noch nicht viel, und darum müssen sie unterrichtet werden und lernen. Dadurch werden sie verständig.

Ich werde in der Schule von Lehrern unterrichtet. Ich bin meinem Lehrer Dankbarkeit und Gehorsam schuldig. So lange ich unterrichtet werde, bin ich ein Schüler. Ein guter Schüler ist aufmerksam; er hört nur auf das, was der Lehrer sagt, und denkt nur an das, was er thun, oder begreifen und behalten soll.


Von der Erde und ihren Bewohnern.

Daß die Erde sehr groß, aber doch nur ein kleiner Theil der Welt sei, haben wir schon gehört. Was für eine Gestalt die Erde habe, ist schwer auszumachen, weil man nur einen sehr kleinen Theil der Erde auf ein Mal übersehen kann, und weil sie uns zu nahe ist. — Aus dem Schatten eines Körpers kann man mit ziemlicher Gewißheit erkennen, ob er rund, breit, oder eckig und spitzig sei; und wenn der Schatten eines Körpers von allen Seiten alle Mal, so oft er sich zeigt, rund erscheint, so ist nicht zu zweifeln, daß auch der Körper rund sei. Dies ist nun der Fall bei unserer Erde. Ihr habt wohl schon von Mondfinsternissen gehört? Bei diesen erblickt man in der Mondscheibe alle Mal einen runden Schatten, und es ist ausgemacht, daß dieser Schatten von unserer Erde in den Mond geworfen wird, so oft sie bei ihrem Um-

Mehrschrittiger Workflow

Kurze Sắtze zur Erweckung der Aufmerkſam-
keit und des Nachdenkens.

C&
ı gehőre zu den Kindern. Kinder wiſſen noch nicht
viel, und darum műſſen ſie unterrịchtet werden und
lernen. Dadurch werden ſie verſtảndia.
Ich werde in der Schule von Lehrern unterrichtet. Ich
bin meinem Lehrer Dankbarkeit und Gehorſam ſchuldig.
So lange ich unterrichtet werde, bin ich ein Schűler.
Ein guter Schűler iſt aufmerkſam; er hőrt nur auf
das, was der Lehrer ſagt, und denkt nur an das, was er
thun, oder begreifen und behalten ſoll.

IV.
Von der Erde und ihren Bewohnern

Daß die Erde ſehr groß, aber doch nur ein kleiner Then
der Welt ſei, haben wiv ſchon gehőrt. Was fűr eine Ge-
ſtalt die Erde habe, iſt ſchwer auszumachen, weil man nur
einen. ſehr kleinen Theil der Erde auf ein Mal őberſehen
kann, und weil ſie uns, zu nahe iſt. — Aus dem Schatten
eines Kőtbers kann man mit ziemticher Gewißheit erkennen,
ob er rund, breit, oder ekkig und ſpitzig ſei; And wenn der
Schatten eines Kőrpers von allen Seiten alle Mal, ſo oft er
ſich zeigt, rund erſcheint, ſa iſt nicht zu zweifeln, daß auch der
Kőrper rund ſei. Dies iſt nun der Fall bei unſerer Erde. Jyr
habt wohl ſchon von Mondſinterniſſen gehőrt? Bei dieſen
erblickt man in der Mondſcheibe ålle Mal einen runden Schat-
ten, und es iſt ausgemacht, daß dieſer Schatten von unſerer
Erde in den Mond geworfen wird, ſo vft ſie bei ihrem Am-

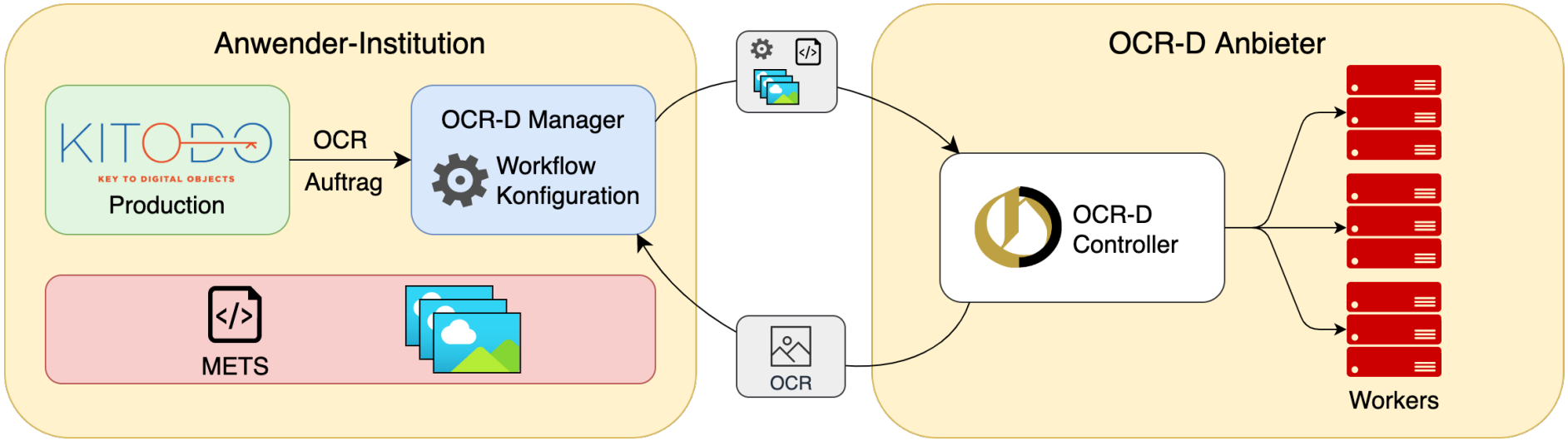
Einschrittiger Workflow

.Kurze Sắtze zur Erweckung der Aufmerkſam-
keit und des Nachdenkens.

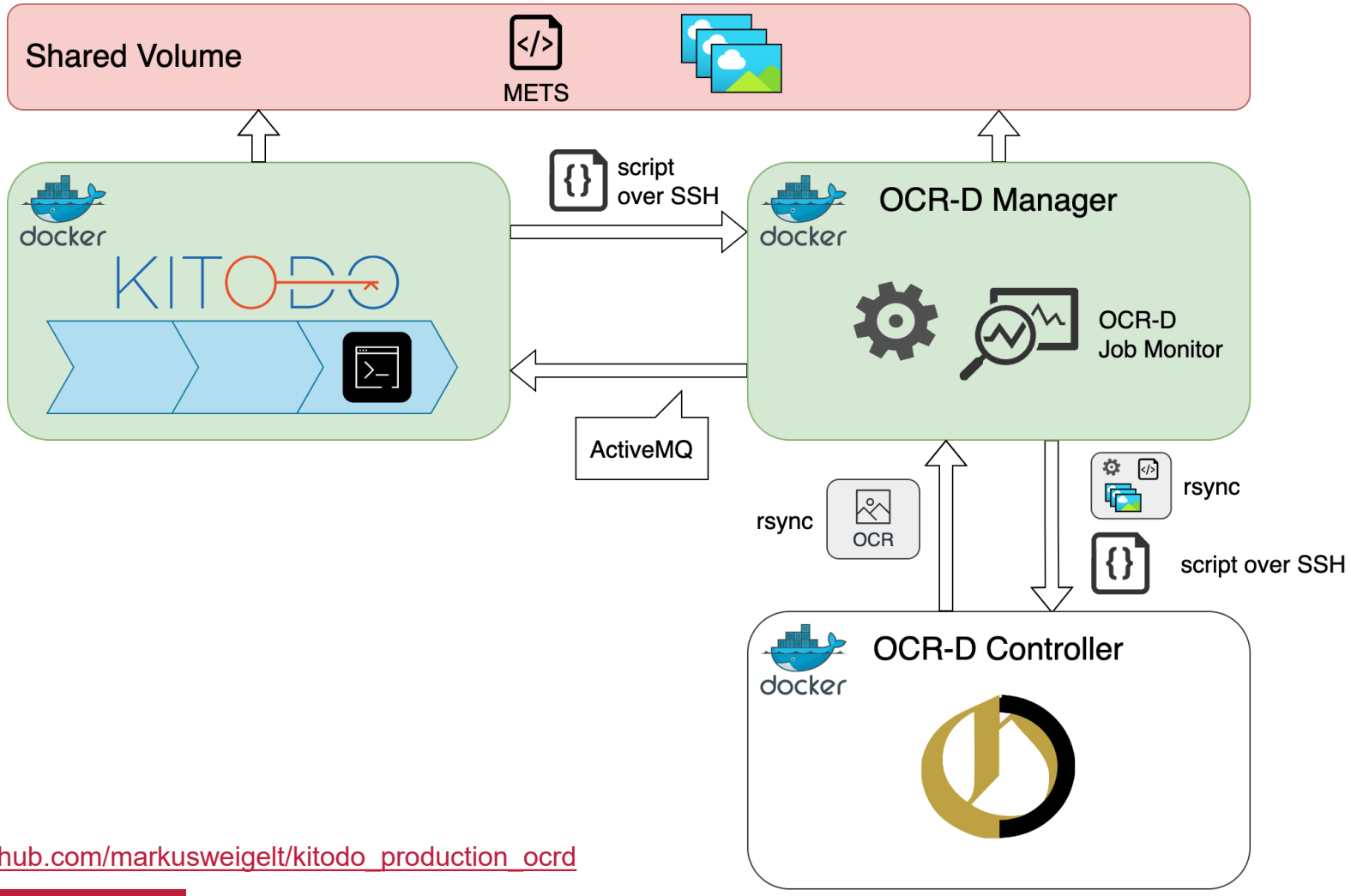
I-oa gehőre zu den Kindern. Kinder wiſſen noch nicht
vi el, und darum műſſen ſie unterrichtet werben ind
ler nen. Dadurch werden ſie verſtảndig.
Ich werde in der Schule von Lehrern unerrichtet. Ich
bin meinem Lehrer Dankbarkeit und Gehorſam ſchuldig.
So lange ich unterrichtet werde, bin ich ein Saler.
Ein guter Schűler iſt a ufmerkſam; er hőrt u auf
das, was der Lehrer ſagt, und denkt nur an daß, baß
thun/ oder begreifen und behalten ſoll.

22 J n da Ghe hid ihte J —
de 298 —— 90 Wa m —9 — An — 120 12
ſtat — rde habe J ſhwer auszumachen, weil 58 — 3 der Erde auf ein — 9 00
—— — — Seiten 7 — 9 9 er ſe Gelſt — 5 A
K B run eh henfuft c —— —
erblickt wohl: in 69 von Rod Ad Nal Mmen —
Erde in den Mond geworfen wird, ſo — —

Geplante Implementierung



Prototyp

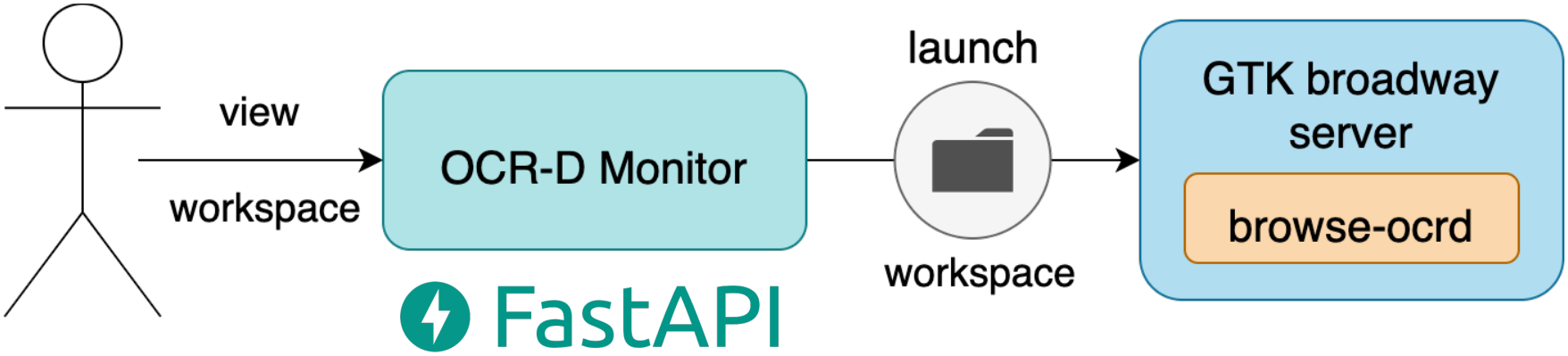


https://github.com/markusweigelt/kitodo_production_ocrd

OCR-D Monitor

Teil des OCR-D Managers

- Status von OCR-D Jobs überwachen
- Mit Ergebnissen interagieren



Weitere Projektschritte

- Fertigstellung der automatisierten Kitodo-OCR-D Verbindung
- Optimierung der Workflowauswahl (Geschw.-Qualität Verhältnis!)
- MyCore-DFG Viewer Verknüpfung (seitens OCR-D)

Danke für Ihre Aufmerksamkeit!