

Digitalisierung & OCR

USB Coffee Lecture, 01.06.2023

Christoph Bartmann, USB Köln
bartmann@ub.uni-koeln.de

Motivation

- Retrodigitalisierung als “Kerngeschäft” der USB Köln, die den größten Altbestand in Nordrhein-Westfalen verwaltet
- Grundlegende Überarbeitung des Digitalisierungs-Workflows und der Präsentations-Systeme und Formate

aber auch:

- Retrodigitalisierung als “Experimentierfeld” für neue technische Möglichkeiten und Standards.
- “Experimentierfeld” auch für Aufgaben und Ziele der Bibliothek im 21. Jahrhundert

Ziele der Bibliothek im 21. Jahrhundert

- Möglichst digitale Verfügbarkeit und Durchsuchbarkeit aller(!) Bestände der Bibliothek im Volltext, bei möglichst hoher Erschließungsqualität
- Volltextzugang und Open Access
- Offene Daten und Datenaustausch durch interoperable Standards
- Möglichst uniformer Zugang zu allen digitalisierten Materialien,
- Einbindung der Nutzenden im Sinne einer (kollaborativen) Arbeit an und mit digitalen Objekten - Ermöglichung von "User Generated Content", "Open oder Citizen Science"

Begrenzte Möglichkeiten für die Mehrzahl der lizenzierten Publikationen

- Entstehung zugriffsgeschützter und geschlossener Verlags-Plattformen statt offener und interoperabler Repositorien
- begrenzte und zweckgebundene Zugänglichkeit wissenschaftlicher Volltexte
- begrenzte und teure Open Access-Angebote

Aber:

Im Rahmen der Retrodigitalisierung eröffnen sich neue Möglichkeiten für die Bereitstellung, Präsentation und vielfältige Nutzung von digitalen Objekten

Retrodigitalisierung als Experimentierfeld neuer Dienste

Kollaboration und Partizipation der User; “User Generated Content”, nutzerspezifische Sammlungen, Annotationen

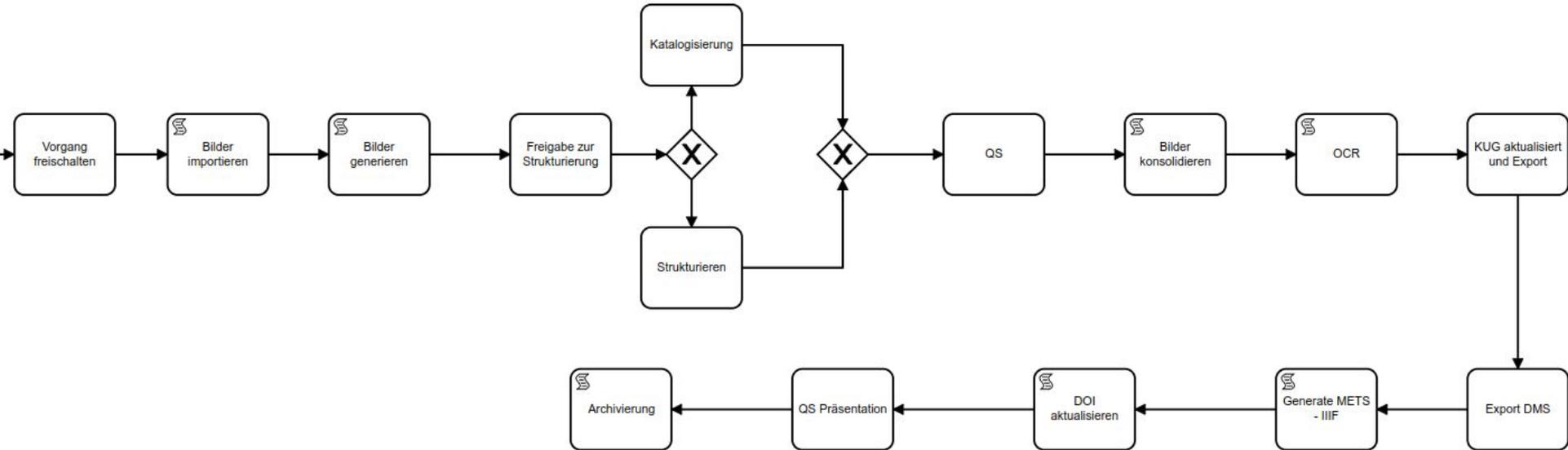
Interoperabilität und Nachnutzbarkeit der veröffentlichten Materialien, Recherchierbarkeit aller Materialien in gemeinsamen Portal, “Neue, systematische Wege” in “Altes Material”

Bereitstellung von Digitalen Objekten und Volltexten im Rahmen der Massendigitalisierung physischer Bestände

Digitalisierungs-Workflow an der USB Köln

- Digitalisierungs-Workflow, der mehrstufigen Prozess implementiert, mit dem Ziel
 - der Massendigitalisierung mit hohem Durchsatz (mehrere Hunderttausend Scan-Seiten pro Jahr)
 - Katalog- und Strukturdaten-Erschließung aller digitalisierten Materialien
 - Möglichst gute Volltexterkennung auf digitalisiertem Material
 - Bereitstellung aller digitalen Objekte über standardisierte und offene Schnittstellen
 - Präsentation aller Digitalisate, Volltexte und Sammlungen in einem gemeinsamen Sammlungsportal

Digitalisierungs-Workflow an der USB Köln



Digitalisierungs-Workflow mit Kitodo Production



Aufgaben



Alle Aufgaben



Titel	Vorgang	Projekt	Aktionen
Freigabe zur Strukturi...	retro_423484_050003	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_421464_050002	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_423499_050005	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_137156_050225	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_140141_050301	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_143637_050292	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_154964_050251	USB Retro Monograph...	
Freigabe zur Strukturi...	retro_160022_050258	USB Retro Monograph...	

Vorgänge



Alle Vorgänge



ID	Titel	Status	Korrektur	Projekt	Aktionen
813	retro_5021...			FID Philoso...	
812	retro_4669...	<div><div style="width: 100%;"></div></div>		USB Retro ...	
811	retro_4661...	<div><div style="width: 100%;"></div></div>		USB Retro ...	
810	retro_4659...	<div><div style="width: 100%;"></div></div>		USB Retro ...	
809	retro_4653...	<div><div style="width: 100%;"></div></div>		USB Retro ...	
808	retro_4653...	<div><div style="width: 100%;"></div></div>		USB Retro ...	
807	retro_4651...	<div><div style="width: 100%;"></div></div>		USB Retro ...	
806	retro_4650...	<div><div style="width: 100%;"></div></div>		USB Retro ...	

Digitalisierungs-Workflow mit Kitodo Production

The screenshot displays the Kitodo Production web interface for editing metadata. The top navigation bar includes the logo, a search field, and the current page title 'Metadaten bearbeiten'. The document ID 'retro_4661281_050275 (ID: 811)' is shown, along with action buttons: 'Schließen', 'Validieren', 'Speichern', and 'Speichern & Schließen'.

The interface is divided into three main sections:

- Strukturdaten:** A tree view showing a 'Monographie' folder with 20 sub-items.
- Metadaten:** A form for editing document metadata:
 - Dokumenttyp:** Monographie
 - Katkey:** 4661281
 - Label:** Antwort des erztstift-köln.
 - Jahr der Digitalisierung:** 2023
 - Identifer:** retro_4661281_050275
 - DOI-Präfix:** 10.58016
 - Sprache:** Deutsch
- Galerie:** A grid of 20 thumbnail images representing the scanned pages of the document.

At the bottom, there are checkboxes for 'Kommentare' and 'Paginierung'. The footer contains links for 'Impressum', 'Datenschutz', and 'Nutzungsbedingungen'.

International Image Interoperability Framework

- Internationaler (de facto-) Standard zur Präsentation von Bilddaten und digitalen Objekten im Internet

Ziel:

- Bessere Nutzbarkeit und Interoperabilität der in unterschiedlichen Repositorien vorgehaltenen Bilddaten
- Spezifikation einheitlicher Methoden um auf Bilder und digitale Objekte zuzugreifen, auch über Repositoriumsgrenzen hinweg
- Ermöglicht die einfache Nutzung von Bilddaten in unterschiedlichen Forschungskontexten



statt...

- organisationsspezifische Mechanismen für die Bereitstellung und den Abruf von Bilddaten (und digitalen Objekten), die nur schwer nachnutzbar sind
- Einsatz selbst entwickelter oder proprietärer Präsentations-Technologien

...mit IIIF

- einheitliche URL-Struktur für die Bereitstellung von Bildern und Bild-Segmenten in unterschiedlichen Auflösungen
⇒ IIIF-Image-API
- einheitliche Mechanismen für die Publikation (und damit auch die Präsentation) strukturierter digitaler Objekte
⇒ IIIF-Presentation-API

IIIF-Anwendungen - IIIF Viewer (Mirador)

The image shows a screenshot of the IIIF Viewer (Mirador) interface. The interface is split into two main panes. The left pane displays a self-portrait of Paul Gauguin, titled "Self-Portrait Dedicated to Paul Gauguin". The portrait is set within an ornate, dark wooden frame. Below the main image is a horizontal strip of six small thumbnail images, with the third thumbnail (the current image) highlighted by a red border. The right pane displays a traditional Indian painting, titled "Bodleian Library MS. Ind. Inst. Misc. 22". The painting depicts Krishna and Radha seated face-to-face on a red and yellow platform. Krishna is on the left, with blue skin, wearing a yellow dhoti and a red shawl. Radha is on the right, with fair skin, wearing a red sari with green and black stripes. They are both adorned with jewelry. Krishna is holding a peacock feather in his right hand. The background is a plain, light brownish-tan color. The interface includes a navigation sidebar on the left with icons for home, search, and settings. At the top, there are window management icons (maximize, refresh, close) for each pane. At the bottom right, there is a page indicator: "1 of 84 • [f. 1 Krishna and his lover Radha seated face-to-face]".

Volltexterkennung (OCR)

Ziel: Hochwertige Volltexterkennung möglichst aller Retrodigitalisate, um Volltextsuche innerhalb des gesamten Digitalisats-Bestands zu ermöglichen

- Zunehmend gute Open Source Lösungen für die Volltexterkennung gedruckten historischen Materials
- Volltexterkennung gedruckter Bücher des 20. Jahrhunderts als weitgehend “gelöstes” Problem
- OCR weiterhin aufwendig auf Materialien mit komplexen Layout insbesondere im Kontext der Massendigitalisierung



OCR-D

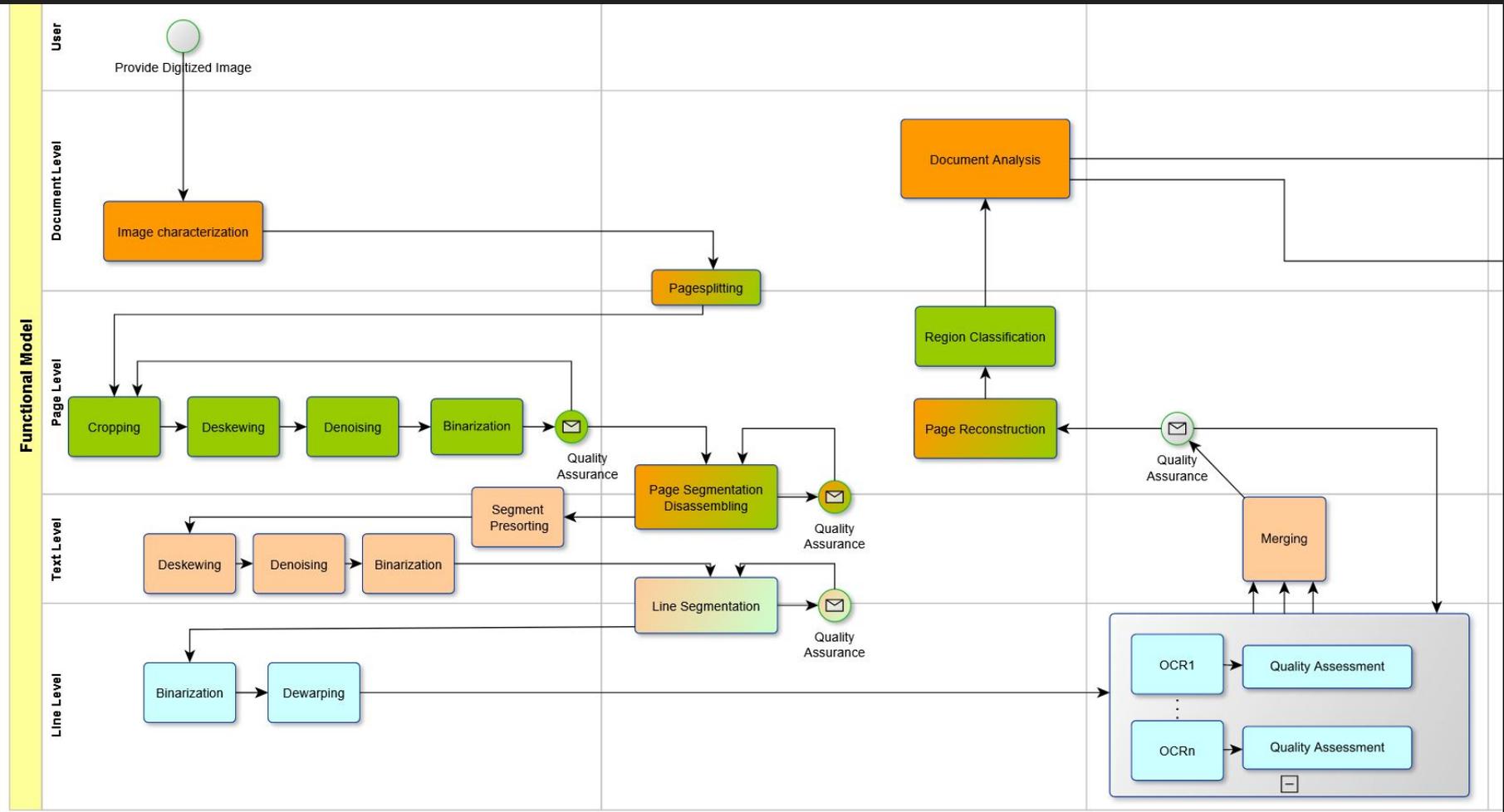
Koordinierte Förderinitiative zur Weiterentwicklung
von Verfahren der Optical Character Recognition (OCR)

Systematischer Einsatz von Open Source Technologien aus dem
OCR-D-Projekt im Rahmen des Digitalisierungsworkflows

USB als ine der ersten Bibliotheken, die Software-Pakete des OCR-D-Projekts
direkt an die Kitodo-Workflow-Software anbindet

Das OCR-D-Projekt als OCR-Workflow-Architektur, die Volltexterkennung als
komplexen Prozess begreift

OCR-D-Workflow

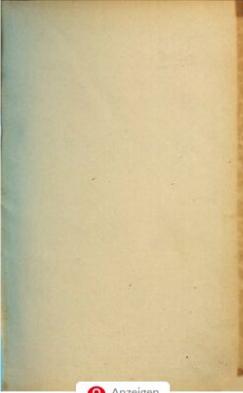


Suche in Metadaten und Volltext - USB-Digitalisate

Bismarck 

Suche in Metadaten Volltexten | Filter aktive Filter beibehalten

< Seite 1 von 19.081 >

 **Otto v. Bismarck** : Von [H. Robolsky] (Bismarck). 3, Bismarck in Frankfurt

Leipzig: Renger, 1885
München, Bayerische Staatsbibliothek -- Biogr. 1652-3/4

[Vollständige Informationen](#)  Permalink 



26 Herr von Bismarck und seine Kollegen. 

24 Herr von Bismarck und seine Kollegen. 

Herr von Bismarck und seine Kollegen. 25 

Herr von Bismarck kommt in späteren Berichten wied 

Vernetzter Wissensraum - USB-Digitalisate



Köln Provenienzportal

Informationen Register Suchhistorie Suchen Suchen Erweiterte Suche | Alle Titel

Start > Informationen > Sammlung



Allgemeine Information zu Sammlungsvermerken

Die in der USB vergebenen Sammlungsvermerke ermöglichen uns, alle Titel einer Sammlung in einem separaten Katalog zusammenzuführen. Diese Kataloge sind Teil des jeweiligen Sammlungsportals. Überdies können wir mittels des Sammlungsvermerks Bestände, die nicht (mehr) zusammen aufgestellt sind, virtuell wieder vereinigen. Eine Übersicht über alle Sammlungsportale finden Sie in den Sammlungen der USB.

Sammlungsvermerke im Kölner Provenienzportal

Die Einträge im Register "USB Sammlungsvermerk" führen zu den Titeln einer Sammlung, die einen Sammlungsvermerk erhalten haben und zusätzlich ein spezifisches anderes Provenienzmerkmal aufweisen. Bei Mehrfachexemplaren kann anhand des Katalogs jedoch nicht genau das Exemplar bestimmt werden, auf das sich der Sammlungsvermerk bezieht. Bei den anderen Provenienzeinträgen ist dies jedoch der Fall, weil sie strukturiert mit Angabe des Exemplars erfasst wurden.

ExLibris Stephan Broelmann (1551 - 1622), Jurist und Historiker aus Köln



DIGITALE EINBANDSAMMLUNG USB KÖLN

Informationen Projekt Register Übersichten Suchhistorie Merklisse Suchen Suchen Erweiterte Suche | Alle Digitalisate

Start > Recherche: **Buchbinder/Werkstätten** Augustiner-Chorherrenstift <Rebdorf bei Eichstätt> / Werkstatt > Titel

RDF JSON

Digitale Einband-Sammlung der USB Köln

Titel-Daten Einband-Daten Abbildungen

Anzahl Abbildungen	1
Beschreibung als PDF	[Download]

	[Vollbild]	Auflösung	2898x2117
		Dateigröße	17,66 MB

Behalten Sie diesen Titel:

- An BibSonomy senden
- In die Merklisse
- Ab speichern
- Ausdrucken

Andere Formate

Literaturliste

Tagging