ZOP
Zurich Open Platform

Pascal Becker, Stephan Schmid

Vom Scanner direkt in DSpace

Die Publikation von Retrodigitalisaten auf der Zurich Open Platform

DSpace Praxistreffen 31.03.2022

The Library Code GmbH

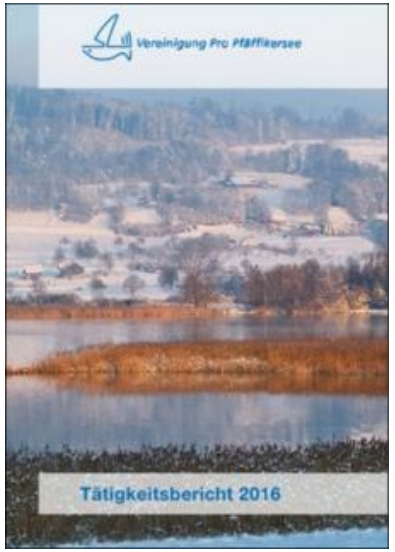
ZB Zentralbibliothek Zürich

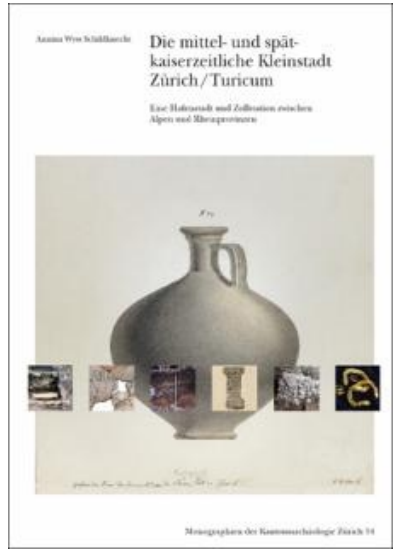
Zentralbibliothek Zürich

- Öffentliche Stiftung
- Entstanden 1917 aus dem Zusammenschluss der Stadt- und Kantonsbibliothek Zürich
- Stadt-, Kantons- und Universitätsbibliothek von Zürich
- Sammelauftrag für Turicensia gemäß Statuten

Turicensia

- Die Sammlung Turicensia nach 1800 umfasst Titel zum Thema Zürich sowie Medien von Zürcher Autor:innen und im Kanton Zürich erschienene Publikationen.
- Zuständig für die Sammlung ist die gleichnamige Abteilung.





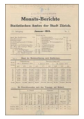
Projekt ZOP

- Handschriften und Alte Drucke aus den Beständen Schweizer Bibliotheken (u.a. auch der ZB) haben mit e-manuscripta und e-rara ihre eigenen Plattformen
- Born digital sowie Retrodigitalisate nach 1800 finden auf diesen Plattformen keinen Platz
- Ziel I: Wahrnehmung des kantonalen Sammelauftrags auch für digitale Neuerscheinungen (Born digital) und wichtige digitalisierte Bestände des 19. und 20. Jahrhunderts
- Ziel II: Ein Open Access Repositorium für (thematische) Turicensia nach 1800
- Aktuell: ca. 1.000 Dokumente online, weitere in Vorbereitung

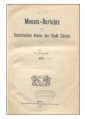
ZOP
Zurich Open Platform

Suche

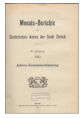
Aktuellste Veröffentlichungen



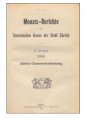
BUCH
Monats-Berichte des Statistischen Amtes der Stadt Zürich 1913



BUCH
Monats-Berichte des Statistischen Amtes der Stadt Zürich 1911



BUCH
Monats-Berichte des Statistischen Amtes der Stadt Zürich 1910

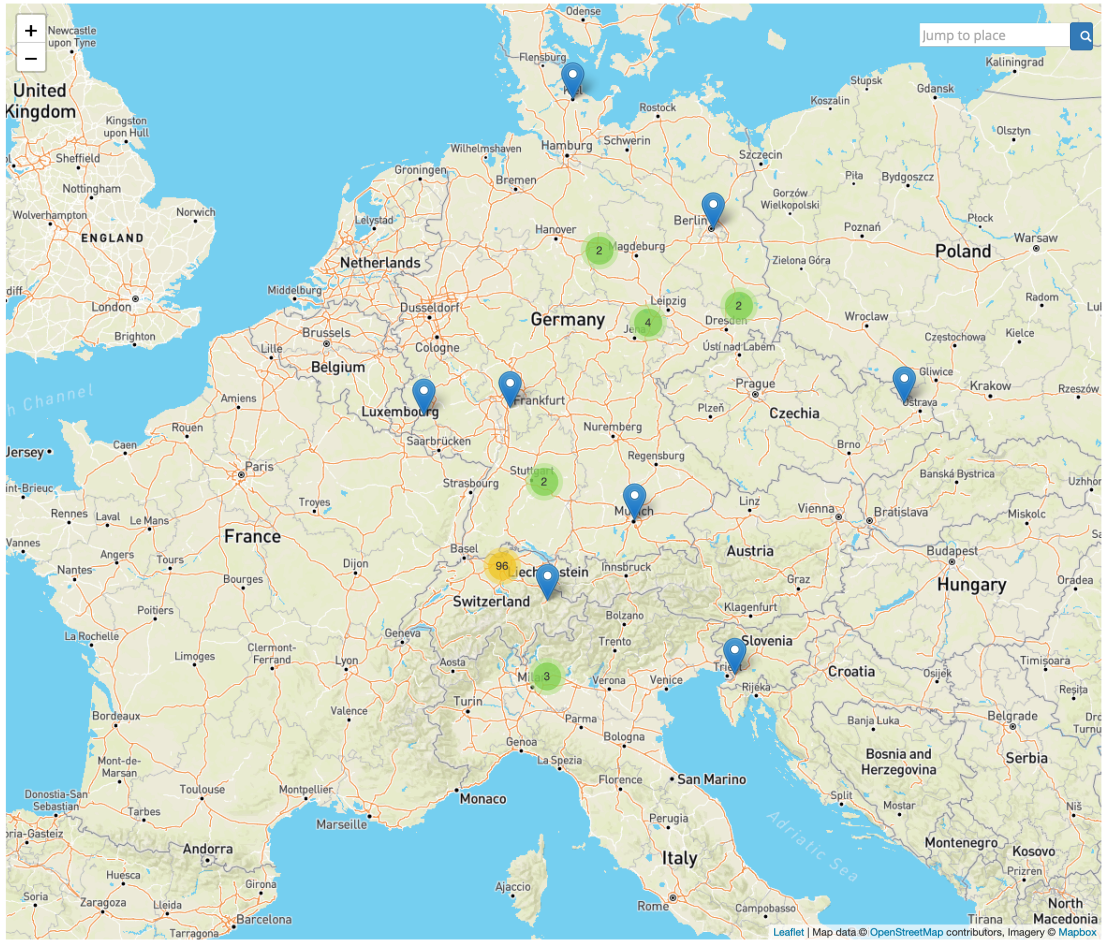


ZEITSCHRIFTENHEFT
Monats-Berichte des Statistischen Amtes der Stadt Zürich 1908

ZEITSCHRIFTENHEFT
Gemeindefinanz-Statistik für das Jahr 1921

ZEITSCHRIFTENHEFT
Gemeindefinanz-Statistik für das Jahr 1913

Zurich Open Platform



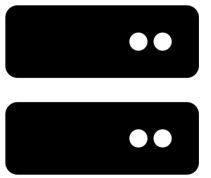
Zurich Open Platform

- Repositorium auf Basis von DSpace 6, JSPUI
- Thematische Turicensia
  (Informationen, die inhaltlich in Bezug zu Stadt und Kanton Zürich stehen)
- Erweiterungen im Vergleich zu DSpace 6:
  - GND-Vergabe
  - Anzeige von Karten für GND-Geographika
  - PDF-Vorschau
  - Metadatenimport aus dem Bibliothekskatalog zu Beginn der Eingabe möglich (via SRU)
- Meilenstein: Go-Live von ZOP Juni 2021: https://zop.zb.uzh.ch
- Partner: Kantonale Archäologie und Denkmalpflege, Zürcher Hochschule der Künste, Amt für Städtebau der Stadt Zürich

Workflow I

phys. Bestände aus Archiv werden im Digitalisierungszentrum gescannt

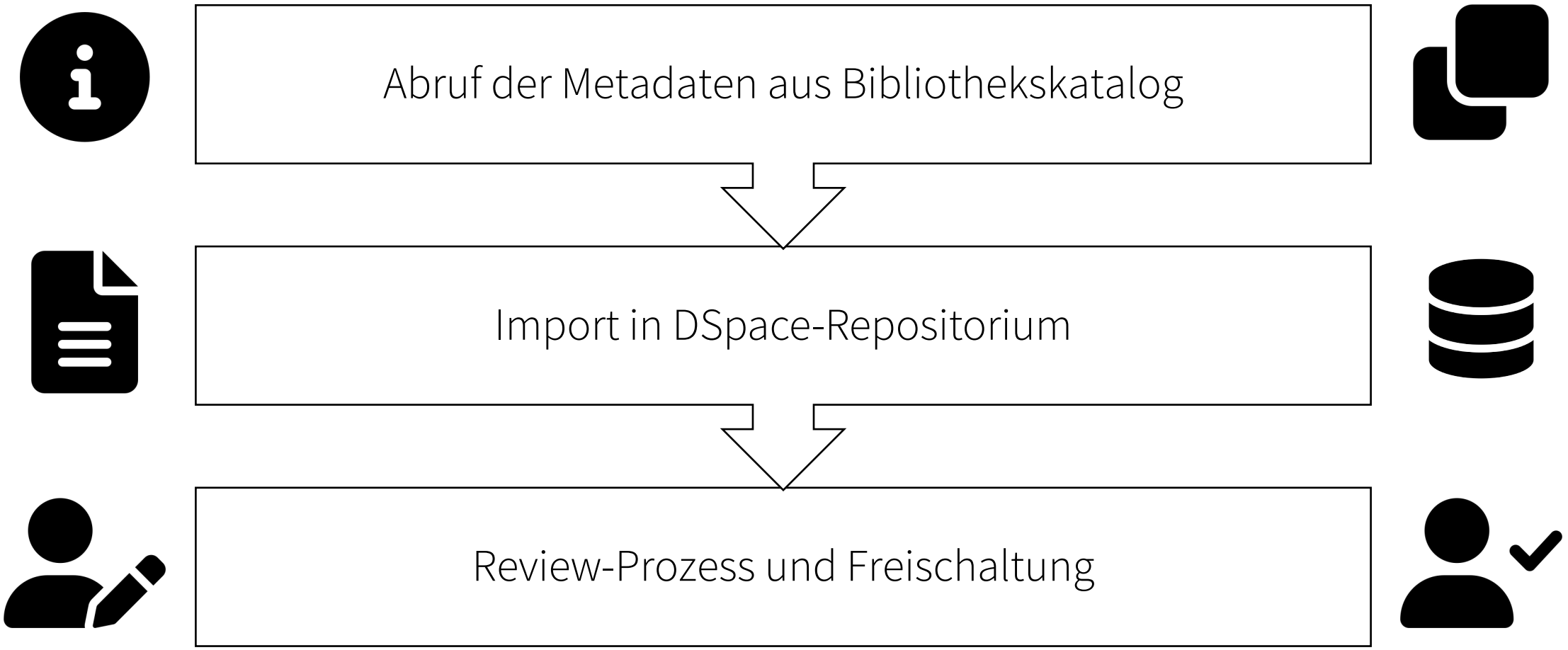
TIFF-Dateien in Ordner gesammelt und auf TLC-Server hochgeladen



OCR-Erkennung und Zusammenführung der TIFFs in PDF

Icons: Font Awesome 5, CC-By: https://fontawesome.com/license

Workflow II



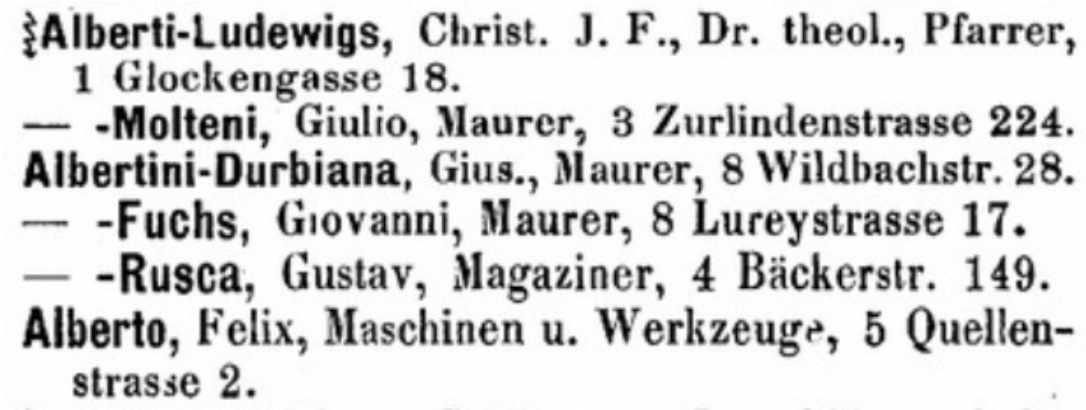
Icons: Font Awesome 5, CC-By: https://fontawesome.com/license

Optical Character Recognition (OCR) 

- Open-Source OCR-Software Tesseract (Apache 2.0 License)
- Voraussetzung: TIFF-Dateien mit mindestens 300 dpi
- Wichtige Information: Handelt es sich um Antiqua- oder Fraktur-Schrift?
- Nutzung eines OCR-Trainingsmodell für Frakturschrift des Projektes OCR-D
  - https://ocr-d.de/
  - https://ub-backup.bib.uni-mannheim.de/~stweil/ocrd-train/data/Fraktur_5000000/tessdata_fast/Fraktur_50000000.334_450937.traineddata
- Erzeugt hOCR
  - standardisiertes Format
  - Enthält Metadaten, erkannten Text und Informationen über dessen Position im Bild

Icon: Font Awesome 5, CC-By: https://fontawesome.com/license

OCR: Antiqua

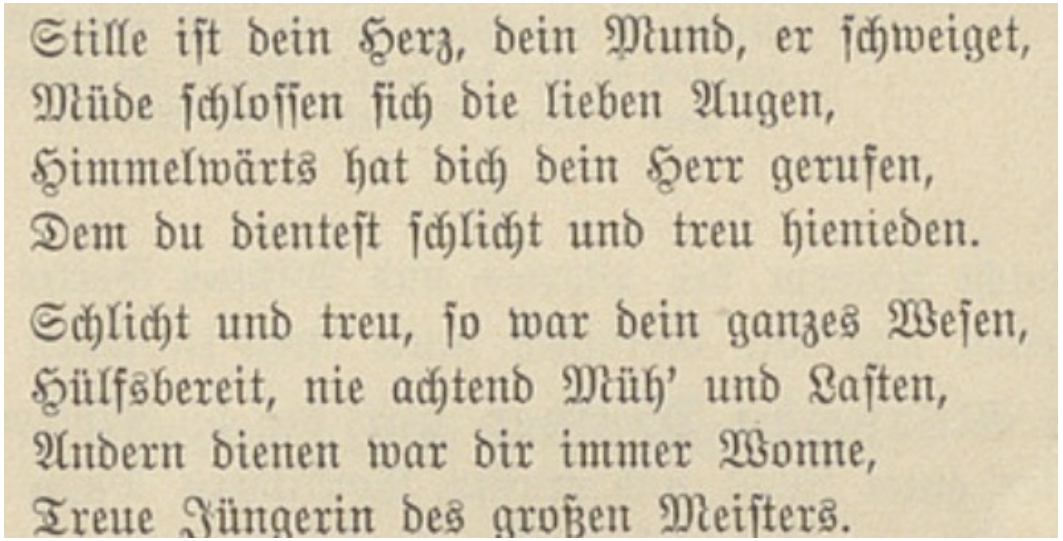


Alberti-Ludewigs, Christ. J. F., Dr. theol., Pfarrer, 1 Glockengasse 18.
— **-Molteni**, Giulio, Maurer, 3 Zurlindenstrasse 224.
Albertini-Durbiana, Gius., Maurer, 8 Wildbachstr. 28.
— **-Fuchs**, Giovanni, Maurer, 8 Lureystrasse 17.
— **-Rusca**, Gustav, Magaziner, 4 Bäckerstr. 149.
Alberto, Felix, Maschinen u. Werkzeuge, 5 Quellen-strasse 2.

- !Alberti-Ludewigs,Christ.J. F., Dr. theol.,Pfarrer,1 Glockengasse18.— -Molteni,Giulio,Maurer,3 Zurlindenstrasse224.Albertini-Durbiana,Gius.,Maurer,8 Wildbachsir.28.— -Fuchs,Giovanni,Maurer,8 Lureystrasse17.— -Rusca,Gustav,Magaziner,4 Bäckerstr.149.Alberto,Felix,Maschinen u. Werkzeuge,5 Quellen-strasse2.!

Adressbuch der Stadt Zürich 1921 (https://doi.org/10.20384/zop-210)

OCR: Fraktur



Stille iſt dein Herz, dein Mund, er ſchweiget,
Müde ſchloſſen ſich die lieben Augen,
Himmelwärts hat dich dein Herr gerufen,
Dem du dienteſt ſchlicht und treu hienieden.

Schlicht und treu, ſo war dein ganzes Weſen,
Hülfsbereit, nie achtend Müh' und Laſten,
Andern dienen war dir immer Wonne,
Treue Jüngerin des großen Meiſters.

- Stilleist deinHerz,deinMund,erschweiget,Müde schlossensich die liebenAugen,Himmelwärtshat dichdeinHerrgerufen,Demdudientests chlichtundtreuhienieden.Schlichtundt reu,so wardeinganzesWesen,Hülfsbereit,nie achtendMüh'undLasten,Anderndienen wardir immerWonne,TreueJüngerindes großenMeisters.

Anna Elisabetha Däniker (Nekrolog, https://doi.org/10.20384/zop-166)

PDF-Dateien

- Jede Seite wird als eine TIFF-Datei gescannt
- Hohe Auflösung gewünscht und auch für OCR erforderlich
- Unkomprimierte TIFF-Dateien in 300 dpi sind groß
- Zur Anzeige besser geeignet: PDF-Dateien
  - Eine Datei je Band
  - Erkannter Text wird „unsichtbar“ vor das Bild gelegt
  - Volltext-Suche innerhalb eines PDFs
  - Mit imagemagick konvertiert, transformiert und zusammengesetzt
- Kompromiss zwischen Dateigrößen (Ladezeiten, Speicher) und Auflösung: 150 dpi

Abruf der Metadaten aus Bibliothekskatalog

- Beim Ablegen auf unseren Servern wird ein Ordner je gescanntem Werk angelegt
- Ordnername enthält MMS-ID des Bibliothekkatalogs
- Via SRU werden Metadaten aus dem Bibliothekskatalog zur MMS-ID abgerufen
- Wandlung der Metadaten aus MARC in die genutzten Metadatenschemata des Repositoriums
- Wird vom Script automatisiert genutzt, kann auch manuell zu Beginn eines neuen Eintrags genutzt werden
- Die MMS-ID wird in den Metadaten gespeichert
- Über OAI-PMH können die Digitalisate später in die Discoervy-Lösung des Bibliothek eingebunden werden

Import in DSpace

- Metadaten und PDF-Dateien werden in DSpace importiert
- Importe erfolgen als abgeschlossene Submissions in den Review-Prozess
- Manuelle Ergänzungen und gegebenenfalls Korrekturen im Rahmen der DSpace-Workflows
- OCR, PDF-Erstellung und Import erfolgen vollautomatisch

Nachnutzung

- Der Quellcode steht unter der DSpace Source Code License (BSD-Lizenz)
- Der Quellcode ist allerdings speziell auf ZOP und die Zentralbibliothek Zürich ausgerichtet (insbesondere XSLT für die Transformation der Metadaten) und müsste erweitert/angepasst werden
- Bestandteile:
  - Erweiterung von DSpace für den Import von Metadaten via SRU (Kommandozeile und für die Maske zur Aufnahme neuer Inhalte ins Repositorium)
  - Python-Script zur Abwicklung von OCR und PDF-Konvertierung
  - Python-Script zum Import in DSpace
- Bei Interesse: Bitte senden Sie uns eine E-Mail (contact@the-library-code.de)

Live-Demonstration