Masterarbeit Tagebuch

Sven Burkhardt

Inhaltsverzeichnis

---

1 Vorlage dd.mm.yyyy

Kurzbeschreibung

Hier kurze Zusammenfassung einfügen.

Erledigte Aufgaben

Aufgabentitel

✓ Aufgabenbeschreibung.
Ergebnis: Ergebnisbeschreibung.

Nächste Schritte

❑ Beschrieb der nächsten Aufgabe.

Offene Fragen

❑ Beschreibung der offenen Frage.

2 Tagging der JPGEGs im AppleFinder 26.10.24

Kurzbeschreibung

Überlegung: JPEGs sollen bereits im Apple Finder mit Tags versehen werden, um eine effiziente, automatisierte Transkription der Chorunterlagen des Männerchors Murg zu ermöglichen. Geplant ist die Kombination von ChatGPT und Transkribus zur Erkennung unterschiedlicher Dokumententypen. Ein Tag-System, bestehend aus **„Maschinell“** für maschinengeschriebene und **„Handschrift“** für handschriftliche Dokumente, gewährleistet die gezielte Zuordnung zur jeweils geeigneten OCR-Software *(Maschinenschrift mit ChatGPT, Handschrift mit Transkribus "German Giant").*

Dokumente, die sowohl maschinell erstellten Text als auch handschriftliche Elemente enthalten, werden entsprechend ihrer Hauptinformationsgehalt getaggt. Zusätzlich erhalten alle Dokumente mit Unterschriften den Tag **„Unterschrift“**, um eine gezielte Verarbeitung dieser Elemente sicherzustellen.

2.1 AppleFinder Tags

- **Handschrift**
- **Maschinell**
- **mitUnterschrift**
- **Bild**

Erledigte Aufgaben

Handschriften tagging

✓ taggen **Handschrift** in AppleFinder.
✓ taggen **Maschinell** in AppleFinder.
✓ taggen **Bild**
✓ taggen **mitUnterschrift**

Ergebnis: Handschriften, Maschinell, Bilder und alle handschriftlichen Unterschriften getaggt

Nächste Schritte

❑ Skripte schreiben, um maschinelle Text zu extrahieren ❑ Transkribus für Handschriftliches anschmeißen.
❑ Nach Gemeinsamkeiten in den Texten suchen, um automatisierte Abfrage für ChatGPT zu erstellen.
❑ Ggf. Aufteilung in unterschiedliche Korpora (Briefe handschr. Briefe schreibmaschine, Zeitungsunterlagen.)
❑ Transkribus für Handschriften verwenden.

Offene Fragen

❑ Sollen die Bilder gelöscht werden?

3 JPG Datenbereinigung - leere Seiten löschen 25.10.2024

JPG Datenbereinigung

Alle JPGs ohne Inhalt, also beispielsweise Rückseiten, werden gelöscht. Regel: sobald etwas handschriftlich oder gedruckt auf einer Seite steht, bleibt es erhalten. Im Moment sind auch Bilder (Bsp. Postkarten inbegriffen.Bilder mit Taggs

3.1 Anmerkung

Geschichte/Chronik/Gründung des Männerchors in Akte 323

Erledigte Aufgaben

JPG Datenbereinigung

✓Alle JPGs ohne Inhalt, also beispielsweise Rückseiten, werden gelöscht.
Ergebnis: Reiner JPG Korpus mit Schriftgut, aber auch Bildern (bspw. Postkarten)

Nächste Schritte

❑ Nach Gemeinsamkeiten in den Texten suchen, um automatisierte Abfrage für ChatGPT zu erstellen.
❑Ggf. Aufteilung in unterschiedliche Korpora (Briefe handschr. Briefe schreibmaschine, Zeitungsunterlagen.)
❑ Transkribus für Handschriften verwenden.

Offene Fragen

❑ Sollen die Bilder gelöscht werden?
❑ Handschriftliche, maschinengeschriebene und gemischte Daten taggen? Ggf. erst später mit ChatGPT.
❑ Transkribus für Handschriften verwenden?

---

4 Datennormalisierung PDF zu JPEG 24.10.2024

Kurzbeschrieb

Heute zwei Python-Skripte zur Normalisierung der Akten geschrieben:

Erledigte Augaben

PDF zu JPG Konvertierung

✓ Skript ***JPEG-to-PDF.py*** geschrieben.
Ergebnis Alle PDF-Seiten in JPGs umgewandelt, Dateinamen mit Seitenzahlen formatiert.

Prüfung der Aktennummern

✓ Skript ***Check-if-all-files-complete.py*** geschrieben.
Ergebnis: Überprüft, ob Akten von 001 bis 425 vorhanden sind. Alle Akten sind vollständig in JPG umgewandelt.

Nächste Schritte

❑ Daten für OCR-Bereinigung vorbereiten, leere Seiten manuell entfernen.

Offene Fragen

❑ Handschriftliche, maschinengeschriebene und gemischte Daten taggen? Ggf. erst später mit ChatGPT.
