



Digitale Harmonie aus historischer Dissonanz

Extraktion, Ordnung und Analyse
unstrukturierter Archivdaten
des Männerchor Murg

Sven Burkhardt

 0009-0001-4954-4426

 17-056-912

 15.08.2025




University
of Basel



Digital
Humanities
Lab

University of Basel
Digital Humanities Lab
Switzerland



Abstract

Diese Arbeit befasst sich mit dem Archiv des Männerchor Murg in den Jahren des Zweiten Weltkrieges. Hierfür wird eine automatisierte Pipeline auf Basis von LLMs und Pattern-matching vorgestellt, mit deren Hilfe Named Entities extrahiert und weiterverarbeitet werden. Ziel ist es, dieses Archiv digital zugänglich zu machen, die beteiligten Personen sowie deren Netzwerke und dessen geographische Ausdehnung sichtbar zu machen.

Inhaltsverzeichnis

Einleitung

Ziel und Relevanz der Arbeit

Formulierung der Forschungsfrage

Aufbau der Arbeit

Geografischer und historischer Kontext

Die vorliegende Arbeit beschäftigt sich mit Unterlagen aus dem Archiv des "Männerchor Murg" und dessen Nachfolger, den "New Gospelsingers Murg". Murg ist eine Gemeinde am Hochrhein, rund 30 km Luftlinie von Basel entfernt. Der Ort liegt am gleichnamigen Fluss Murg, der in den Rhein mündet. Beide Gewässer bildeten über Jahrhunderte hinweg den wirtschaftlichen Motor der Region: Die Wasserkraft der Murg begünstigte früh die Ansiedlung von Mühlen, Hammerwerken und Schmieden entlang des Bachlaufs. Parallel bot der Rhein mit seiner Drahtseil-Fähre eine wichtige Verkehrs- und Handelsverbindung, und bis zum Ersten Weltkrieg privat betrieben wurde.

Mit dem Ausbau der Landstraße, der heutigen Bundesstraße 34, sowie dem Anschluss an die Bahnstrecke Basel–Konstanz wandelte sich Murg im 19. Jahrhundert von einer landwirtschaftlich geprägten Siedlung zu einer Gewerbe-, Handels- und Industriegemeinde. Während der Industrialisierung Mitte des 19. Jahrhunderts wurde die Wasserkraft zu einem entscheidenden Faktor der Region. Die Ansiedlung der Schweizer Textilfirma *Hüssy & Künzli AG* im Jahr 1853¹ trug wesentlich zum Wachstum der Gemeinde bei. Zahlreiche Arbeitskräfte — auch aus der benachbarten Schweiz — machten die Gemeinde zu einem wichtigen Standort der regionalen Textilindustrie. Es waren eben diese schweizer Textilarbeiter, die 1861 den *Männerchor Murg* gründeten, damals unter dem Namen *Schweizer Männerchor Murg*.

1. Vgl. Gemeinde Murg, Hrsg., *Geschichte Gemeinde Murg*,
besucht am 29. Juni 2025, <https://www.murg.de/seite/33378/geschichte.html>.

Quellen

Quellentradierung

In den Lagerräumen der New Gospel Singers Murg, dem Nachfolgeverein des Männerchors Murg, wird im Jahr 2018 mehrere je ca. 800 Seiten umfassende Ordner mit historischen Unterlagen gefunden. Für diese Arbeit wird ein Ordner mit der Aufschrift *“Männerchor Akten 1925-1944”* gewählt, da er neben dem Ordner *“Männerchor Akten 1946-1950”* den größten Zeitraum abdeckt. Darüberhinaus bietet er das Potential, aufschlussreiche Einblicke in das Vereinsleben in der Zeit vor und während des Nationalsozialismus, insbesondere des Zweiten Weltkrieges, zu geben.

Der Ordner umfasst insgesamt 780 Seiten und deren Inhalt kann als “Protokoll”, “Brief”, “Postkarte”, “Rechnung”, “Regierungsdokument”, “Noten”, “Zeitungsartikel”, “Liste”, “Notizzettel” oder “Offerte” kategorisiert werden.

Quellenbeschreibung

Zeitraum

blabla

Dokumententyp

blabla

Inhalt???

Korpus

Aus dem Bestand des Ordners *“Männerchor Akten 1925-1944”* werden für diese Arbeit ausschließlich Akten verwendet, die während des Zweiten Weltkriegs verfasst wurden. Der Analysezeitraum erstreckt sich dementsprechend zwischen dem 01. September 1939 und dem 8. Mai 1945², dem Tag der bedingungslosen Kapitulation Deutschlands.

Diese Begrenzung des zeitlichen Rahmens ist notwendig, um die Funktionalität der weiter unten beschriebenen Pipeline darstellen zu können. Hieraus ergibt sich in der Folge

2. [\[\[vgl.\]\]](#)[\[Finde hier eine Referenz\]](#)keylist

eine Limitation der Anzahl potetieller Akteurinnen und Akteure, Orte und Organisationen. Notwendig ist das besonders mit hinblick auf die Erstellung einer verlässlichen Groundtruth mit angereicherten historischen Daten durch Archivrecherchen. Sie helfen, das Potential solch einer digital unterstützten Arbeitsweise zu unterstreichen.

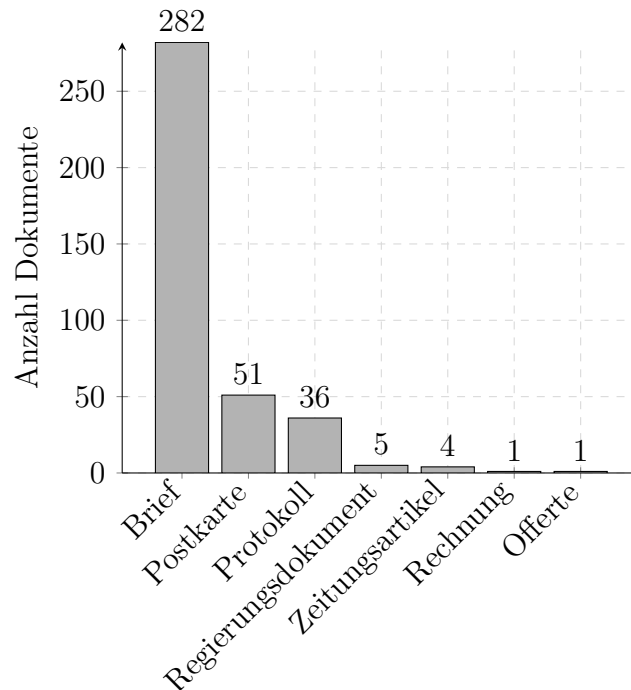


Abbildung 1: Verteilung der Dokumententypen im untersuchten Bestand (150 Akten - 381 Seiten).

Digitalisierung

blabla

Sichtung & Kategorisierung in Akten

blabla

Transkription

blabla

Test mit Tesseract

blabla

Test mit LLM

blabla

Transkribus

blabla

Tagging

blabla

Tagging mit Transkribus

blabla

Tagging mit LLM

blabla

Export

blabla

Forschungsstand und Forschungslücke

Die oben beschriebenen Quellen wurden vor dieser Arbeit bislang nicht wissenschaftlich bearbeitet.

In LOD wird auf Vorprojekte verwiesen. Diese bitte beschreiben (Arcgis und Feldpost)

Methodenkritik

Genutze Tools

Digitale Methoden spielen für die Durchführung dieser Arbeit eine zentrale Rolle. Von der Digitalisierung der Quellen über die Transkription bis hin zur Auswertung durchlaufen die Daten zahlreiche Prozessschritte, die mithilfe von Large Language Models, Deep-Learning-Modellen und anderen digitalen Werkzeugen verarbeitet und visualisiert werden. Die Auswahl der Tools orientierte sich dabei an Kriterien wie Verfügbarkeit (Open Source

vs. proprietär), Kompatibilität, Community-Support, erforderlichem Arbeitsaufwand und selbstverständlich dem konkreten Mehrwert für die Forschungsfragen.

In diesem Kapitel werden sowohl Werkzeuge vorgestellt, die tatsächlich eingesetzt wurden, als auch solche, die sich im Verlauf des Projekts als ungeeignet erwiesen. Transparenz ist hierbei ein wesentlicher Aspekt: Ein großer Teil der Methodik entwickelte sich erst im Forschungsprozess selbst. Da sich Large Language Models rasant weiterentwickeln, ist nicht immer von Beginn an klar, ob ein Tool für den eigenen Anwendungsfall geeignet ist. Um diese Unsicherheiten zu dokumentieren, werden hier auch gescheiterte Versuche dargestellt.

LOD - Linked Open Data

Linked Open Data (LOD) bezeichnet einen dezentral organisierten Ansatz zur Veröffentlichung und Verknüpfung strukturierter Daten im Web. Ziel ist es, Datensätze verschiedener Institutionen und Akteure maschinenlesbar zugänglich zu machen und über standardisierte Formate wie RDF und SPARQL miteinander zu verbinden³. Wesentliches Merkmal der LOD-Cloud ist dabei die Nutzung semantischer Beziehungen, insbesondere Äquivalenzen einzelner Daten. Hierfür wird häufig das Prädikat `owl:sameAs` genutzt, um z.B. mit `:Choir owl:sameAs wd:Q131186` eine eigene Instanz als identisch mit der Wikidata-Entität für einen Chor zu deklarieren. Klassen oder Instanzen können so aus unterschiedlichen Datenquellen eindeutig identifiziert und zusammengeführt werden.

Die OWL Web Ontology Language, entwickelt vom World Wide Web Consortium (W3C), ist damit ein zentrales Werkzeug für die Realisierung von LOD.⁴ Mit ihr lassen sich Ontologien definieren, die Domänen über Klassen, Individuen und deren Relationen formal beschreiben. Sie ermöglichen, logische Schlussfolgerungen zu ziehen, um verteilte Datenbestände zu verknüpfen und maschinenlesbar auszuwerten. Besonders relevant ist dabei `owl:sameAs`, das als Identitätsrelation fungiert: Es deklariert Instanzen, die in unter-

3. Vgl. Emmanouel Garoufallou und María-Antonia Ovalle-Perandones, Hrsg., *Metadata and Semantic Research. 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020. Revised Selected Papers*, Bd. 1355, Communications in Computer and Information Science (Madrid, Spain: Springer Nature Switzerland AG, 2. Dezember 2020

), Preface S. VI & S. 13f, ISBN: 978-3-030-71903-6, besucht am 5. Juli 2025, https://basel.swisscovery.org/discovery/openurl?institution=41SLSP_UBS&vid=41SLSP_UBS:live&doi=10.1007%2F978-3-030-71903-6_30.

4. Vgl. „OWL Web Ontology Language Guide“, unter Mitarb. von Michael K. Smith, Chris Welty und Deborah L. McGuinness, (Zugriff am besucht am 5. Juli 2025)

, besucht am 5. Juli 2025, <https://www.w3.org/TR/owl-guide/>.

schiedlichen Quellen unter verschiedenen URIs⁵ geführt werden, als dasselbe reale Objekt⁶ und ermöglicht so eine präzise Zusammenführung von Informationen — ein Grundpfeiler für die Interoperabilität im Semantic Web. Die OWL-Spezifikation baut auf RDF⁷ auf und erweitert es um zusätzliche Konzepte. Die Sprache liegt in drei Varianten vor⁸, die sich im Grad ihrer Ausdrucksstärke unterscheiden.⁹ Insbesondere OWL DL bietet einen praktikablen Mittelweg zwischen hoher Ausdruckskraft und vollständigem, entscheidbarem Schließen (Reasoning) und ist daher für viele LOD-Anwendungsfälle geeignet.

Trotz ihres Potenzials wird diese Form der Datenverknüpfung bislang jedoch nicht von allen Websites konsequent umgesetzt.¹⁰ Für die technische Umsetzung für diese Arbeit werden zwei zentrale Werkzeuge genutzt: Protégé zur Modellierung der Ontologie und GraphDB für deren Verwaltung und Abfrage.

Protégé

Zur praktischen Modellierung der Ontologie kam *Protégé* zum Einsatz. Protégé ist eine weit verbreitete Open-Source-Software zur Erstellung, Visualisierung und Verwaltung von Ontologien. Die grafische Oberfläche unterstützt eine intuitive Klassendefinition, Relationserstellung und Instanzverwaltung. Mit Hilfe von Plugins können darüber hinaus logische Konsistenzprüfungen durchgeführt und Ontologien direkt im OWL-Format exportiert werden, um sie in LOD-Workflows einzubinden. Die initiale Version der Ontologie für dieses Projekt entstand zuerst im Codeeditor *Visual Studio Code* wurde aber schnell vollständig in Protégé überarbeitet. Damit bildet das Programm die Grundlage für erste Experimente mit Abfragen in SPARQL.

GraphDB

Für die Speicherung und Abfrage der Ontologie wurde *GraphDB* verwendet. GraphDB ist eine spezialisierte RDF-Triplestore-Datenbank, die es ermöglicht, große Mengen an semantisch verknüpften Daten effizient zu verwalten. Mit der integrierten SPARQL-Schnittstelle können Benutzer gezielt nach Instanzen, Klassen und Relationen suchen und komplexe Muster in den Datenbeständen erkennen. Im Rahmen dieser Arbeit diente GraphDB als

5. Abk. URI; Uniform Resource Identifier

6. Vgl. „OWL Guide“, 2.3. Data Aggregation and Privacy.

7. Abk. RDF; Resource Description Framework

8. OWL Lite, OWL DL und OWL Full

9. Vgl. „OWL Guide“, 1.1. The Species of OWL.

10. Vgl. Garoufallou und Ovalle-Perandones, *Metadata and Semantic Research*, S. 14.

Backend, um die in Protégé entwickelte Ontologie zu testen und mit realen Entitäten aus den untersuchten Quellen abzugleichen.

mma-Ontologie

Ein wichtiger Aspekt dieser Arbeit ist die Unstrukturiertheit relevanter Informationen. Aus diesem Grund wurde auf der Basis der oben beschriebenen Semantik begonnen, eine eigene Ontologie zu entwickeln, die die identifizierten Entitäten systematisch erfasst¹¹ Beim Schreiben dieser initialen Ontologie aus rund 2000 Zeilen Code erweist sich schnell ein neues Problem. Die Datengrundlage aus den geschilderten Vorprojekten (siehe ??) ist zu klein, um daraus eine aussagekräftige Netzwerkanalyse zu machen. Hierfür erweisen sich die Unterschiede der Daten zusätzlich als zu gross und damit aufwendig. Der Fokus der Arbeit verschiebt sich dementsprechend von der Ontologieentwicklung auf die Extraktion von Entitäten.



Abbildung 2: Ausschnitt der Turtle-Ontologie.

Der bestehende Datensatz ist zu klein, um eine umfangreiche Ontologie lohnend zu machen. Hinzu kommen externe Quellen, und deren Zugänglichkeit. Zuverlässige Quellen für Informationen über militärische Einheiten und deren Feldpostnummern sind das „Forum der Wehrmacht“¹² und der „Suchdienst des DRK“¹³. In beiden Fällen liegen die Daten jedoch nicht als LOD vor, sondern im Forum als einfache Strings und beim Deutschen

11. Abk. mma; Männerchor Murg MasterArbeit.

12. Vgl. Andreas Altenburger, „Lexikon der Wehrmacht“, (Zugriff am besucht am 15. Januar 2023), besucht am 15. Januar 2023, <https://www.lexikon-der-wehrmacht.de/Gliederungen/Infanteriedivisionen/205ID.htm>.

13. Vgl. „DRK Suchdienst | Suche per Feldpostnummer“, DRK Suchdienst; Suche per Feldpostnummer, unter Mitarb. von Christian Reuter, (Zugriff am besucht am 12. März 2025), besucht am 12. März 2025, <https://vbl.drk-suchdienst.online/Feldpostnummer/FPN.aspx>.

Roten Kreuz als OCR-PDF¹⁴ historischer Suchlisten aus der Nachkriegszeit. Ein manuelles Recherchieren dieser Daten scheint zu diesem Zeitpunkt den Rahmen der Arbeit zu sprengen. Die in diesem Schritt geleistete Vorarbeit beim Sortieren und Klassifizieren von Entitäten, besonders in Verknüpfung mit selbst erstellten Wikidata-Klassen wird in späteren Prozessschritten wieder aufgegriffen¹⁵.

Wikidata

Wikidata Wikidata ist eines der zentralen Repositorien für Linked Open Data, und bietet eine hohe Interoperabilität durch standardisierte URIs, SPARQL-Endpunkte und offene APIs zu den Entitäten. Jede Entität erhält dabei eine eindeutige, persistente URI (z.B. `wd:Q131186` für einen Chor), die in LOD-Szenarien als stabiler Referenzpunkt dient. Neben anderen betonen Martinez & Pereyra Metnik (2024) beispielsweise:

*„Wikidata stands out for its great potential in interoperability and its ability to connect data from various domains.“*¹⁶

Wikidata entspricht, ebenso wie das nachfolgend beschriebene GeoNames, den FAIR-Prinzipien: Die Daten sind **F**indable und **A**ccessible, **I**nteroperable und **R**eusable¹⁷.

Im Rahmen dieser Arbeit dient Wikidata als zentrale externe Referenz, um lokal erhobene Entitäten mit international etablierten Datenobjekten zu verknüpfen und so ihre Interoperabilität sicherzustellen. Die Plattform ermöglicht eine eindeutige Identifizierung sowie die maschinenlesbare Anreicherung um zusätzliche Informationen.

Die praktische Umsetzung zeigt jedoch eine strukturelle Einschränkung: Trotz systematischer Verknüpfung eigens angelegter Datensätze, etwa mit Armeen, Militäreinheiten, Orten und Personen, entfernt die Community-Moderation etwa 70% dieser Beiträge. Dies verweist auf die hohen internen Qualitätsanforderungen, begrenzt jedoch die Verlässlichkeit und den Nutzen der Arbeit erheblich. Aufwand und Unsicherheit über die Persistenz der Einträge machen den ursprünglich vorgesehenen LOD-Ansatz in dieser Form nicht praktikabel.

14. OCR = Optical Character Recognition

15. siehe Kapitel Nodegoat

16. Roxana Martinez und Gonzalo Pereyra Metnik, „Comparative Study of Tools for the Integration of Linked Open Data: Case study with Wikidata Proposal“.

17. wilkinson_fair_2016.

GeoNames

Ebenso wie Wikidata bietet *GeoNames* eine Open-Source-Plattform für interoperable Daten. GeoNames fokussiert sich hierbei auf geografische Informationen und stellt eine umfassende Datenbank mit über 25 Millionen Ortsnamen und rund 12 Millionen eindeutigen geografischen Objekten bereit. Alle Einträge sind in neun Feature-Klassen und über 600 spezifische Feature-Codes kategorisiert. Die Plattform integriert Daten zu Ortsnamen in verschiedenen Sprachen, Höhenlagen, Bevölkerungszahlen und weiteren Attributen aus unterschiedlichen nationalen und internationalen Quellen. Sämtliche Geokoordinaten basieren auf dem WGS84-System¹⁸ und können über frei zugängliche Webservices oder eine API abgerufen werden. Darüber hinaus erlaubt GeoNames registrierten Nutzenden, bestehende Datensätze über eine Wiki-Oberfläche zu bearbeiten oder zu ergänzen, wodurch eine kollaborative Qualitätssicherung gewährleistet wird.

Msty

Alphabet – Gemini

Anthropic – Claude

OpenAI – ChatGPT

Transkribus

Abreviation Tag klappt nicht (wird nicht exportiert) Listen werden nicht exportiert

Nodegoat

Verweis auf Groundtruth in Kombination mit wikidata und geojson, da gementioned in mmma-Ontologie

18. *WGS84: geodätische Grundlage des Global Positioning System (GPS)*, vgl.: „WGS84 | Landesamt für Geoinformation und Landesvermessung Niedersachsen“, Landesamt für Geoinformation und Landesvermessung Niedersachsen, (Zugriff am besucht am 5. Juli 2025), besucht am 5. Juli 2025, https://www.lgln.niedersachsen.de/startseite/wir_uber_uns/hilfe_support/lgln_lexikon/w/wgs84-190576.html.

Netzwerkanalyse als Methode

Theoretischer Hintergrund der Netzwerkanalyse

Ziele der Netzwerkanalyse im Kontext der Quellen

Technische Umsetzung (Tools, Datenbankstruktur)

Pipeline

Aufbau XML to JSON Pipeline

Übersichtsgrafik der Pipeline

Module im Detail

`document__schemas.py`

`__init__.py`

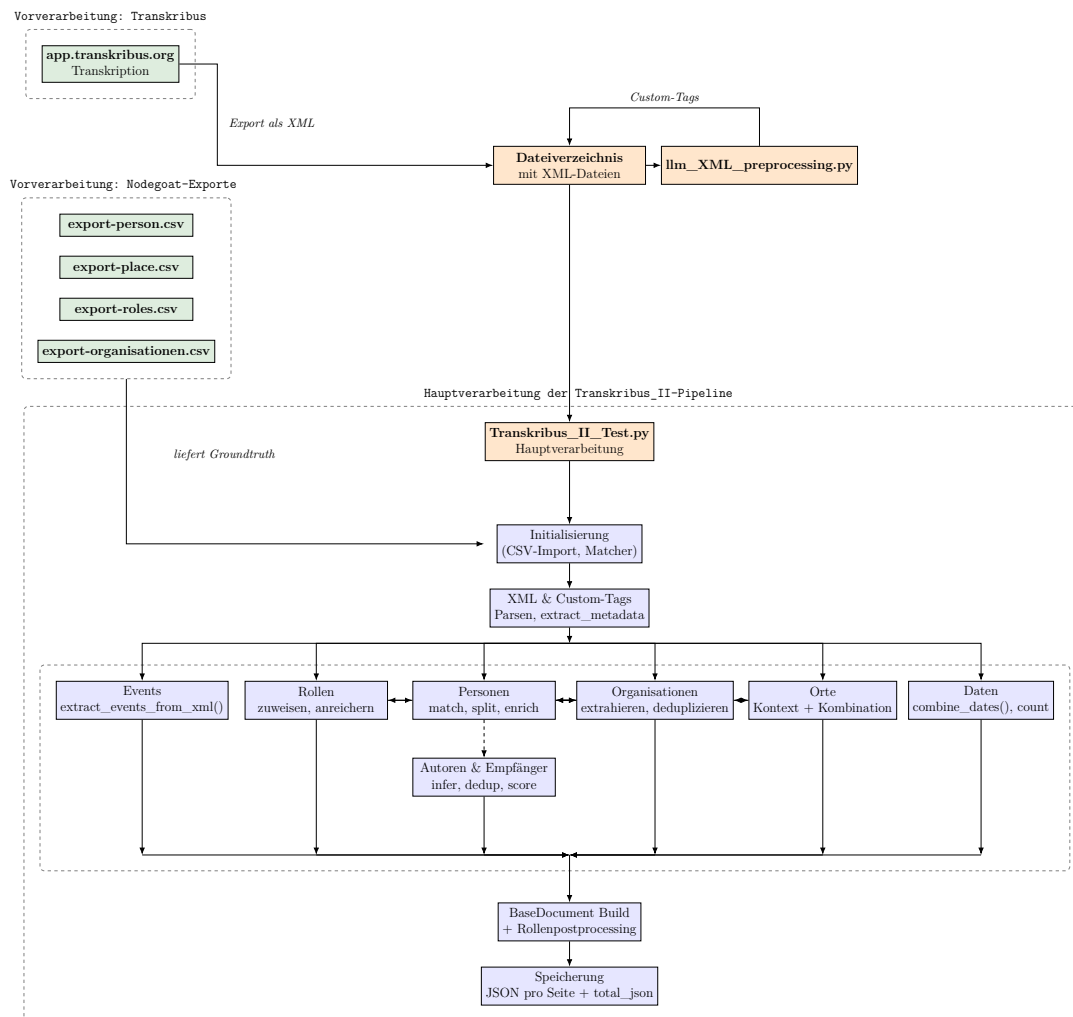


Abbildung 3: Übersicht der gesamten XML-to-JSON-Pipeline

Person-Matcher

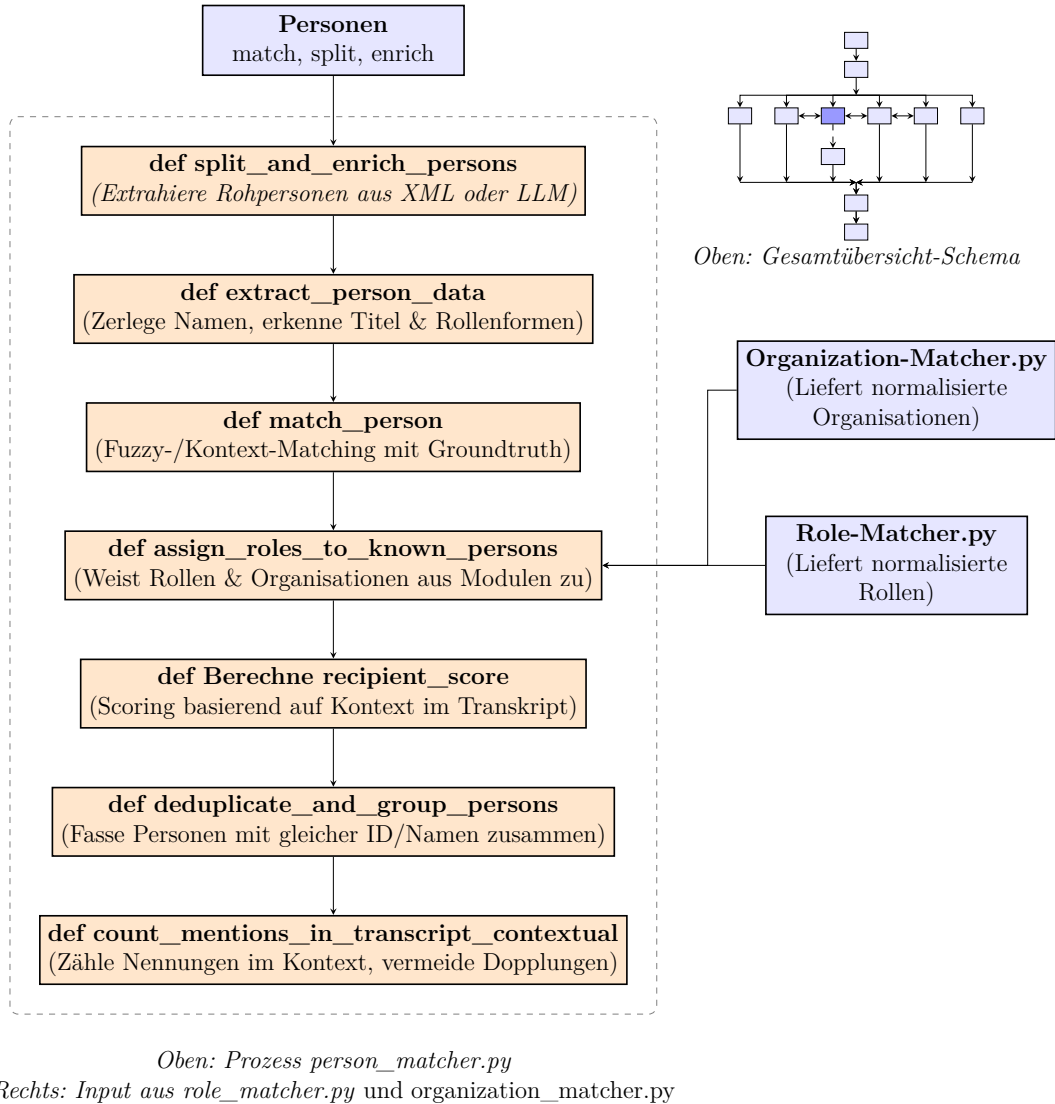


Abbildung 4: Detailliertes Prozessdiagramm: Personen-Matching

place__matcher.py

organization__matcher.py

letter__metadata__matcher.py

type__matcher.py

event__matcher.py

date__matcher.py

Assigned_Roles_Module.py

unmatched__logger.py

KEINE AHNUNG WAS DIE HIER MACHEN

validation__module.py

validation__module.py

test__role__schema.py

llm__enricher.py

enrich__pipeline.py

Analyse & Diskussion der Ergebnisse

Visualisierung auf der VM

Fazit und Ausblick

Zusammenfassung der zentralen Erkenntnisse

Methodische Herausforderungen und Lösungen

Ausblick auf zukünftige Forschung und mögliche Erweiterungen
der Datenbank

ALTER SCHEISS

In Transkribus-Seminaren am Departement Geschichte der Universität Basel wird aus „*Männerchor Akten 1925-1944*“ bereits 2018 und 2022 ein erster Korpus aus 137 Akten¹⁹. Es entsteht eine Liste, die die Seiten mit deren Lage im Ordner, einem Kurztitel und einem Entstehungsdatum versieht. Als Akte werden im Folgenden Schriftstücke bezeichnet, die entweder durch die Fundsituation, oder ihren Inhalt eindeutig als zusammengehörig betrachtet werden können. So liegt Akte__001 beispielsweise in einer separaten Mappe und umfasst 96 Seiten, während andere Akten nur aus einer einzelnen Seite bestehen können. Während der Fokus 2028 auf den augenscheinlich häufig auftretenden Personennamen "Carl Burger" und "Fritz Jung" liegt, wird 2022 im Rahmen eines zweiten Seminars spezifischer die Feldpost untersucht. Zu diesem Zeitpunkt erfolgt die Transkription mit einem generischen Modell, das nicht auf die unterschiedlichen Handschriften trainiert ist.

Forschungsstand zu den Quellen

Die vorliegenden Arbeit stützt sich auf diese Vorarbeit, und die darin gesammelten Daten. Beispielsweise werden die Feldpostbriefe um weitere Daten ergänzt. Kernfragen hierfür sind: *Welche Einheiten verbergen sich hinter den Feldpostnummern? Wo waren die Einheiten, als der Brief geschrieben werden?*

Hierzu werden Nachschlagetabellen in Fachliteratur²⁰, die Bestände des *Bundesarchives – Militärarchiv Freiburg*²¹, des *Suchdienstes des Deutschen Roten Kreuzes (DRK)*²², sowie Citizen Scientist Projekte²³ herangezogen und letztere teils durch eigene Recherche

19. Weiterführend vgl. (Sven Burkhardt, „ArcGIS StoryMaps“, ArcGIS StoryMaps, [Zugriff am besucht am 12. März 2025]

, besucht am 12. März 2025, <https://storymaps.arcgis.com>)

20. vgl.: (Georg Tessin, *Verbände und Truppen der deutschen Wehrmacht und Waffen-SS im Zweiten Weltkrieg 1939-1945*, Bd. Band 1 - Die Waffengattungen — Gesamtübersicht [Osnabrück: HIBLIO Verlag, 1977]

)), (Christian Hartmann, *Wehrmacht im Ostkrieg - Front und militärisches Hinterland 1941/42*, 2. Auflage, Bd. 75, Quellen und Darstellungen zur Zeitgeschichte Herausgegeben vom Institut für Zeitgeschichte [München: R. Oldenbourg Verlag, 2010]

)), (Christoph Rass und René Rohrkamp, *Deutsche Soldaten 1939-1945 Handbuch einer biographischen Datenbank zu Mannschaften und Unteroffizieren von Heer, Luftwaffe und Waffen-SS* [Aachen, 2009]

))

21. Prof. Dr. Michael Hollmann, „Freiburg“, Bundesarchiv Freiburg im Breisgau (Abteilung Militärarchiv), (Zugriff am besucht am 12. März 2025)

, besucht am 12. März 2025, <https://www.bundesarchiv.de/das-bundesarchiv/standorte/freiburg/>.

22. „DRK Suchdienst | Suche per Feldpostnummer“.

23. vgl. Wikidata (Beispiel: („78th Sturm-Division (Wehrmacht)“, unter Mitarb. von Sven Burkhardt, [Zugriff am besucht am 12. März 2025]

, besucht am 12. März 2025, <https://www.wikidata.org/wiki/Q125489568>)),

(„Lexikon der Wehrmacht“, unter Mitarb. von Andreas Altenburger, [Zugriff am besucht am 12. März

ergänzt.

Für diese Arbeit wird die Kategorisierung von 2018 übernommen und auf den Seiten im Ordner erweitert.²⁴

Beschreibung des Archivbestands

2025]

, besucht am 12. März 2025, <http://www.lexikonderwehrmacht.de/>),
(„Forum Geschichte der Wehrmacht“, Forum Geschichte der Wehrmacht, unter Mitarb. von Dieter Hermans, [Zugriff am besucht am 12. März 2025]

, Forum, besucht am 12. März 2025, <https://www.forum-der-wehrmacht.de/>)

24. Akten_Gesamtübersicht.csv im Anhang

Methodischer Zugang

Digitale Erfassung und Strukturierung der Quellen

Gliederung in Akten

Die analogen Akten müssen zuerst für die Digitalisierung vorbereitet werden. Sie werden aus den Ordnern genommen und vorsichtig von Heftklammern, Gummibändern und Büroklammern befreit. Dies dient der Konservierung des Papiers – gerade an Stellen, an denen sich vorher Büroklammern befunden haben, frisst sich Rost in das Papier und beschädigt es stark. Auch sonstiger Säurefraß durch nicht-säurefreies Papier, das sich im Ordner befand, zeigt sich an einigen Stellen.

Um schnell und dennoch in guter Auflösung zu digitalisieren, wird die „Dateien“-App²⁵ von Apple benutzt, da sie gleichzeitig einen grossen Cloud-Speicher und eine OCR-Erkennung bietet. Die Intention dahinter sind schnell durchsuch- und auffindbare Texte. Um die Geschwindigkeit der Digitalisierung zu erhöhen, und eine vergleichbare Qualität zu erhalten, wird ein Ipad mit einem Stativ verwendet, das im 90°Winkel über den Seiten positioniert ist. Die Dateien werden entsprechend der bereits erwähnten Akten_Gesamtübersicht benannt. Sind mehrere Blätter zusammengeheftet, so ergeben sie eine Akte. Sind sie einzeln, werden sie ebenfalls als einzelne Akte geführt. Die Archivierung findet sowohl analog wie digital auf Seiten-Ebene statt.

Digitalisierung und Transkription

Tagging in Transkribus

Transkribus und seine Modelle unterstützen nicht nur beim Transkribieren der Texte, sondern erlauben auch das Taggen von *Named Entities*. Für die vorliegende Arbeit sind dabei besonders Personen, Orte, Organisationen und Daten relevant. Um hierfür ein stringentes Verfahren zu entwickeln, wurden die Tags wie folgt definiert:

25. vgl. [Apple-Finder](#)

abbrev

Mit dem Tag **abbrev** werden alle Abkürzungen getaggt, die für eine eindeutige Entität stehen.

☞ **Beispiel 1:** Dr., Prof., St., Hr., Frl., Dipl.-Ing., etc.

☞ **Beispiel 2:** Organisationskürzel, wenn sie eindeutig sind:

```
<abbrev>V.D.A.</abbrev> .
```

☞ **Beispiel 3:** Falls eine dazugehörige Entität vorhanden ist, wird die Abkürzung getaggt und wird gleichzeitig als zugehörige Entität getaggt:

```
<person><abbrev>Dr.</abbrev>Weiß</person>
```

unclear

Mit dem Tag **unclear** werden unleserliche oder schwer entzifferbare Textstellen markiert.

☞ **Beispiel 1:** Unklare Zeichen oder fehlende Buchstaben:

```
„Er wohnte in<unclear>[...]<unclear>“.
```

☞ **Beispiel 2:** Teilweise lesbare Wörter:

```
"<place>Frei<unclear>[...]<unclear><place>“.
```

sic

Mit dem Tag **sic** werden Wörter markiert, die im Originaltext in einer falschen oder ungewöhnlichen Schreibweise geschrieben wurden.

☞ **Beispiel 1:** Veraltete oder falsche Schreibweisen:

```
„<sic>daß</sic>“ für dass.
```

☞ **Beispiel 2:** Offensichtliche Tippfehler, wenn sie im Originaltext so vorkommen:

```
„Wir haben <sic>einen</sic> große Freude.“
```

☞ **Beispiel 3:** Falls eine Korrektur notwendig ist, kann sie als Kommentar ergänzt werden.

Inhaltliche Tags

person

Mit dem Tag **person** sollen alle Strings getaggt, die eine direkte Zuordnung einer Person ermöglichen.

☞ **Beispiel 1:** Vereinsführer, Alfons, Zimmermann, Alfons Zimmermann, Z. A. Zimmermann, Herr Zimmermann, Herr Alfons Zimmermann, etc.

☞ **Beispiel 2:** Funktionen wie Oberlehrer, Chorleiter, etc., wenn Ort, Name oder Organisation bekannt.

Eine Person kann sowohl mit ihrem Namen als auch ihrer Funktion (wie Dirigent) getaggt werden. Aus der Korrespondenz ist in der Regel eine zugehörige Organisation ersichtlich, mit deren Verknüpfung eine namentlich nicht genannte Person identifiziert werden könnte.

signature

Mit dem Tag **signature** werden alle Strings getaggt, die eine handschriftliche Unterschrift darstellen. Der Tag **signature** ist nahezu deckungsgleich mit dem Tag **person**. Er dient zur **graduellen Unterscheidung**, ob ein Name im Fließtext als gesichert leserlich oder handschriftlich als Signatur vorliegt.

☞ **Beispiel 1:** Eindeutig lesbare Signaturen werden direkt getaggt:

```
<signature>A. Zimmermann</signature>.
```

☞ **Beispiel 2:** Teilweise unleserliche Signaturen werden mit dem Tag **unclear** innerhalb von **signature** markiert:

```
<signature>R. We<unclear>[...]</unclear></signature>.
```

☞ **Beispiel 3:** Wenn nur ein Teil des Namens lesbar ist, aber eine Identifikation unsicher bleibt, sollte die Unterschrift vollständig im Tag **unclear** innerhalb von **signature** stehen:

```
<signature><unclear>Unleserlich</unclear></signature>.
```

☞ **Beispiel 4:** Wenn eine Signatur einer bekannten Person zugeordnet werden

kann, aber nicht vollständig lesbar ist, bleibt die Signatur erhalten und wird **ohne** den Tag **person** zu verwenden:

```
<signature>A. Zimm<unclear>[...]</unclear></signature>.
```

☞ **Beispiel 5:** Wenn eine Unterschrift vollständig transkribiert wurde und die Person bekannt ist, wird sie nur mit **signature** getaggt, **ohne** den Tag **person** zu verwenden:

```
<signature>Alfons Zimmermann</signature>.
```

organization

Mit dem Tag **organization** werden alle Strings getaggt, die eine direkte Zuordnung einer Organisation ermöglichen.

☞ **Beispiel 1:** Männerchor Murg, Verein Deutscher Arbeiter (V.D.A.), Murgtalschule, etc.

☞ **Beispiel 2:** Abkürzungen, wenn sie eine Organisation eindeutig bezeichnen, z.B. V.D.A., NSDAP, STAGMA, etc.

place

Mit dem Tag **place** werden alle Strings getaggt, die sich auf einen geografischen Ort beziehen.

☞ **Beispiel 1:** Murg (Baden), Freiburg, Berlin, Murgtal, Schwarzwald, etc.

☞ **Beispiel 2:** Orte mit näherer Bestimmung, z.B. „bei Berlin“, „im Murgtal“ werden getaggt:

```
<place>im Murgtal</place>.
```

date

Mit dem Tag **date** werden alle expliziten und implizierten Datumsangaben markiert.

☞ **Beispiel 1:** 29.05.1936

☞ **Beispiel 2:** 29. Mai 1936

☞ **Beispiel 3:** den 29. d. Mts.:

```
<date when="29.05.1936 ">den 2.</date> <abbrev>d. Mts.</abbrev>
```

event

Mit dem Tag **event** werden expliziten und implizierten Ereignisse markiert. Diese Ereignisse haben einen zeitlichen oder räumlichen Bezug, und können benannt werden. Dazu zählen:

- ☞ **Beispiel 1:** "Jubiläumskonzert"
- ☞ **Beispiel 2** "Gründung des Vereins"
- ☞ **Beispiel 2** "Kriegsausbruch" oder "Kriegsende"

Konzepte, die nicht klar in den Texten benannt werden, wie beispielsweise die Suche nach einem Dirigenten, können nicht immer Ereignis getaggt werden. Sie sollen später aber in der Datenbank implementiert werden.

Strukturelle Tags

abbrev

Mit dem Tag **abbrev** werden alle Abkürzungen getaggt, die für eine eindeutige Entität stehen.

- ☞ **Beispiel 1:** Dr., Prof., St., Hr., Frl., Dipl.-Ing., etc.
- ☞ **Beispiel 2:** Organisationskürzel, wenn sie eindeutig sind:

```
<abbrev>V.D.A.</abbrev> .
```

- ☞ **Beispiel 3:** Falls eine dazugehörige Entität vorhanden ist, wird die Abkürzung getaggt und wird gleichzeitig als zugehörige Entität getaggt:

```
<person><abbrev>Dr.</abbrev>Weiß</person>
```

unclear

Mit dem Tag **unclear** werden unleserliche oder schwer entzifferbare Textstellen markiert.

- ☞ **Beispiel 1:** Unklare Zeichen oder fehlende Buchstaben:

```
„Er wohnte in<unclear>[...]<unclear>“.
```

- ☞ **Beispiel 2:** Teilweise lesbare Wörter:

"<place>Frei<unclear>[...]<unclear><place>“.

sic

Mit dem Tag `sic` werden Wörter markiert, die im Originaltext in einer falschen oder ungewöhnlichen Schreibweise geschrieben wurden.

☞ Beispiel 1: Veraltete oder falsche Schreibweisen:

„<sic>daß</sic>“ für dass.

☞ Beispiel 2: Offensichtliche Tippfehler, wenn sie im Originaltext so vorkommen:

„Wir haben <sic>einen</sic> große Freude.“

☞ Beispiel 3: Falls eine Korrektur notwendig ist, kann sie als Kommentar ergänzt werden.

Digitalisierungsprozess und Herausforderungen

Hier gehört dringend dazu, dass die Quellen über einen längeren Zeitraum digitalisiert wurden. Das bedeutet, dass sich die Kameras geändert haben. Verwendet wurden primär ein iPad Pro 2nd Generation (2017) und ein iPad Air 4th Generation (2022). Die Verwendete Software ist die Scan-Funktion von Apple iCloud. Die Auswahl der Software war aus rein ökonomischen Gründen. Da das Digitalisierungsprojekt bereits 2018 begonnen wurde, fehlten weitestgehend Grundlagenkenntnisse, die im Digital Humanities Studium vermittelt wurden. Berücksichtigt wurden jedoch einige Richtlinien, wie sie in den Archiv-Kursen des Bachelor-Geschichtsstudiums vermittelt wurden (gleichbleibende Beleuchtung, Hintergrund). Die Scanqualität ist daher oft nicht optimal, was zu Problemen bei der OCR Erkennung mit OCR Software (Apple OCR, Adobe, etc.) führte. Aus diesem Grund wurden 75 Akten zunächst mit dem Model "The German Giant I" mit einer CER von 8,30% transkribiert. Mit insgesamt 4 Iterationen wurde eine Groundtruth für ein eigenes Modell erstellt, und gleichzeitig Personen, Orte, Daten und Organisationen getaggt. Hierzu wurde auch manuell OpenAIs CHatGPT 4o Modell verwendet, das für die Rechtschreibprüfung verwendet wurde. Tauchte ein Rechtschreibfehler im Text auf, wurde dieser manuell überprüft. War der Fehler bereits im Ursprungstext, so wurde der Tag "sic" verwendet, und eine Korrektur beigelegt.

Die so erstellten 70 Akten ergaben 158 Seiten zu insgesamt 22.155 Wörtern Groundtruth, womit dann ein eigenes Transkribus Modell²⁶ (ModelID: 287793) erstellt wurde. Es erreichte eine Accuracy (CER) von 6,58%. Später wurden die verbleibenden 80 Akten nur noch mit diesem Modell transkribiert.

ChatGPT produziert daraus:

Durch CHatGPT verliert der Text zwar seine ursprüngliche Formatierung und Zeilenumbrüche, aber wird nun nahezu fehlerfrei lesbar. Nur das "Venstadler Liedchen" ist eigentlich eines aus "Neustadt". Eine anschließende menschliche Korrektur ermöglicht also den Abgleich mit dem nun lesbaren Text, und die Korrektur der Transkription.

Korrigiert und getagt lautet der Brief nun:

München, 28.V.1941

Lieber Otto!

Nur wer die Sehnsucht kennt weiß was ich leide

Ich wandle traurig her in schwarzer Seide.

Die Sehnsucht brennt, du bist so fern.

Ach lieber Otto, wie hab ich dich gern.

Ich schnitt es gern in alle Rinden.

Ach Otto, wann u. wo kann ich dich finden?

Deine dich nie vergessende

Lina Fingerdick

An

Herrn Otto Bollinger

z.Hd. Herrn Alfons Zimmermann

Vereinsführer des Männerchor

Murg

Laufenburg (Baden)

Rhina

26. burkhardt_transkribus_nodate.

Aug. 15. Aug. 41.

Mein lieber Hans!

Esan lunga brist alung Sane Kaimu-
chor scindes nimmul nien Lindesun zu piffen.
und kum mir die gaffige Gelangensfick zu piffen.

Es ist sehr angenehm, dass Sie uns den
Maimacher Kreisstadt und den T. d. L. Lindgen
zu besuchen, wo Sie zum Abend
essen „auf Hindenburg“ Osnabrück in Paris
mehren beifügt, keine Achtung. Willst
gelingen ab der ersten Zeit zu besuchen.

Drifted fire figures of Sub Line now
" Sub Line now being Kirsch

Es wurde 1928 um 10. Pictus. Längst. Post
um Langstreckungsbahn in die g-fürger
und wutete in einem großen Briffall.

Griff hieser der Richtigkeit zu finden.

Mrs. Wm. : zuruf der Keistader Lidyen.
 u. dem der Wiener Lidyen und vom
 Liden unmöglich, dem sein Wall.

Wm. Lloyd Garrison

Chin

Carl

Abbildung 5: Beispiel für handschriftlichen Text in Akte_076 erkannt mit Transkribus

Murg. 15. Aug 41

Mein lieber Alfons!
Sehen lunge Leitt es mich dem Männer-
chor wieder einmal ein Liedchen zu stehen.
und kam mir die gestege Gelegenheit gussend.
Männechor Venstad um den Title das Liedchen
zu erhalten, wo sie zum Abschied am Aute
sängen „auf Wiederschen Owohl ich Frei!
märke beifügte, keine Aentwarb. Vielleicht
gelingt es Dir diesen litel zu erhalten.
Weiterhin sänge ich fal Lied nur
"Bas alte Lied von being. Rerohl
Es wurde 1928 am 10. Dachub. Sängerb. Frst
von Begrüßungsabend in Dien gesungen.
und erntete überaus großen Reifall.
Es ich schwer das Richtige zu finden.
Aler Alfon, werst das Vemsladler Liedchen.
alsdann das Biener Lidchen und wenn
Leides unmöglich, dann freu Nall.
Mit herzl. Grüße
Dein
Carl

Abbildung 6: Transkription von ??

Murg, 15. Aug. 41

Mein lieber Alfons!

Schon lange treibt es mich, dem Männerchor wieder einmal ein Liedchen zu stiften, und kam mir die günstige Gelegenheit gelegen.

Ich schrieb vergangenes Jahr an den Männerchor Venstad, um den Titel des Liedchens zu erhalten, das sie zum Abschied am Auto sangen: „Auf Wiedersehen, o wohl ich frei!“

Ich fügte eine Frankierung bei, erhielt jedoch keine Antwort. Vielleicht gelingt es Dir, diesen Titel zu erhalten.

Weiterhin sang ich das Lied nur „Das alte Lied von Wien“. Obwohl es am 10. Dezember 1928 beim Sängerbund-Fest von Begrüßungsabend in Wien gesungen wurde und überaus großen Beifall erntete, ist es schwer, das Richtige zu finden.

Aber Alfons, zuerst das Venstadler Liedchen, dann das Wiener Liedchen und wenn beides unmöglich, dann Fröhlichsein.

Mit herzlichen Grüßen

Dein

Carl

Abbildung 7: Transkription durch ChatGPT von ??

Murg. 15. Aug 41

Mein lieber Alfons!

Seit langem treibt es mich dem Männer-chor wieder einmal ein Liedchen zu stiften. und kam mir die günstige Gelegenheit passend.

Ich schrieb vergangenes Jahr an den Männechor Vorstand um den Titel das Liedchen zu erhalten, wo sie zum Abschied am Auto sangen "auf Wiederschen" Obwohl ich Frank-marke beifügte, keine Antwort. Vielleicht gelingt es Dir diesen Titel zu erhalten.

Weiterhin sänge ich das Lied nur

"Das alte Lied" von Komp. Kirchl

Es wurde 1928 am 10. Deutsch. Sängerb. Fest am Begrüßungsabend in Wien gesungen.

und erntete überaus großen Beifall.

Es ist schwer das Richtige zu finden.

Also Alfons! zuerst das Neustadter Liedchen.

alsdann das Wiener Liedchen und wenn

Beides unmöglich, dann freie Wahl.

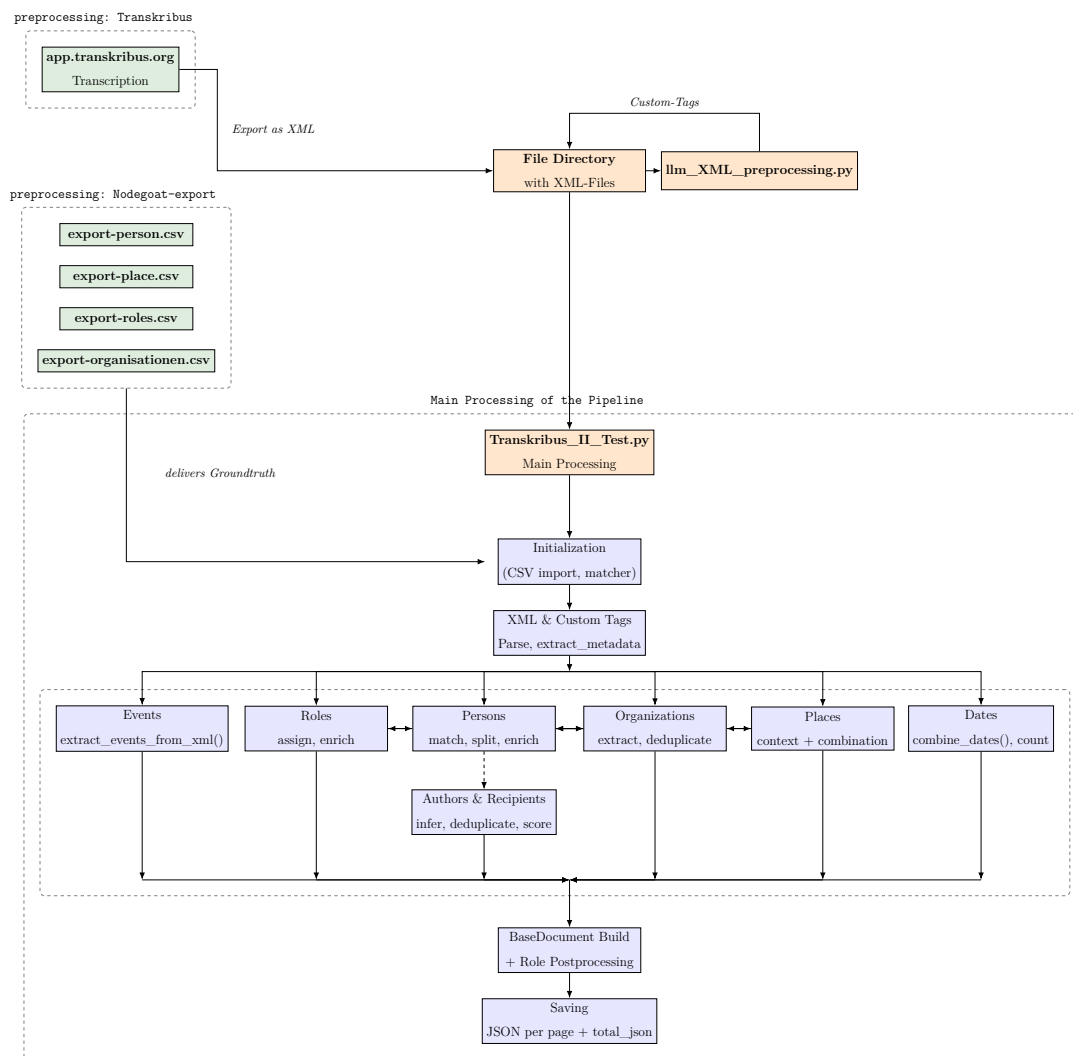
Mit herzlichen Grüßen

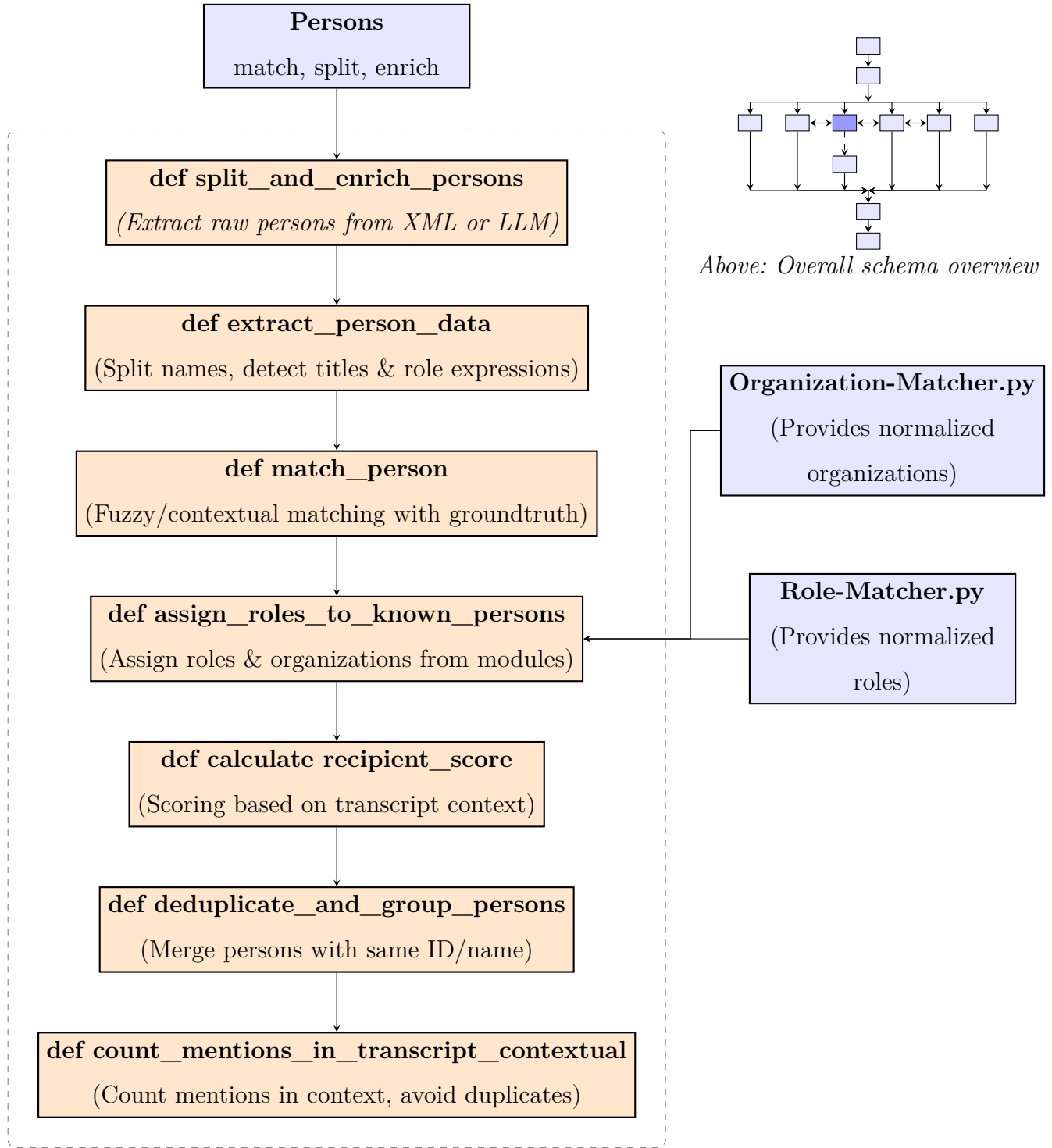
Dein

Carl

Abbildung 8: Tagging von ??

Diagram english





Top: process in person_matcher.py

Right: input from role_matcher.py and organization_matcher.py

Gründe für den Wechsel zu Nodegoat

Nodegoat Modellierung

Netzwerkanalyse als Methode

Theoretischer Hintergrund der Netzwerkanalyse

Ziele der Netzwerkanalyse im Kontext der Quellen

Technische Umsetzung (Tools, Datenbankstruktur)

Normalisierung der Dateien — von PDF zu JPEG

```
1 import os
2 import fitz  # PyMuPDF
3
4 def convert_pdf_to_jpg(src_folder, dest_folder):
5     # Überprüfen, ob der Zielordner existiert, und ihn ggf. erstellen
6     if not os.path.exists(dest_folder):
7         os.makedirs(dest_folder)
8
9     # Durchgehen durch alle Dateien im Quellordner
10    for root, dirs, files in os.walk(src_folder):
11        for file in files:
12            # Überprüfen, ob die Datei eine PDF-Datei ist
13            if file.lower().endswith(".pdf"):
14                # Vollständigen Pfad zur PDF-Datei erstellen
15                pdf_path = os.path.join(root, file)
16                # PDF-Datei öffnen
17                doc = fitz.open(pdf_path)
18                # Durch alle Seiten der PDF-Datei gehen
19                for page_num in range(len(doc)):
20                    page = doc[page_num]
21                    # Seite in ein PiXMap-Objekt umwandeln (für die Konvertierung in
22                    ↪ JPG)
23                    pix = page.get_pixmap()
24                    # Dateinamen ohne Dateiendung extrahieren
25                    filename_without_extension = os.path.splitext(file)[0]
26                    # Ausgabedateinamen erstellen mit führenden Nullen für die
27                    # Seitennummer
28                    output_filename = f"{filename_without_extension}_S{page_num +
29                    ↪ 1:03d}.jpg"
30
31                # Vollständigen Pfad zur Ausgabedatei erstellen
32                output_path = os.path.join(dest_folder, output_filename)
33                # Bild speichern
```



```

33         pix.save(output_path)
34         # PDF-Datei schließen
35         doc.close()
36
37         # Erfolgsmeldung ausgeben
38         print(f"{file} wurde erfolgreich umgewandelt und gespeichert
39         in {dest_folder}")
40
41     # Pfade zu den Ordnern mit den PDF-Dateien (Quelle) und den JPG-Dateien (Ziel)
42     src_folder = r"/Users/svenburkhardt/Documents/D_Murger_Männer_Chor_Forschung/Scan_Mä_
↪ nnerchor/Männerchor_Akten_1925-1945/Scan_Männerchor_PDF"
43     dest_folder = r"/Users/svenburkhardt/Documents/D_Murger_Männer_Chor_Forschung/Master_
↪ arbeit/JPEG_Akten_Scans"
44
45
46     # Funktion aufrufen, um die Konvertierung durchzuführen
47     convert_pdf_to_jpg(src_folder, dest_folder)
48

```

Aufbau der Datenbank

Konzeption der Datenmodellierung

Eigene Ontologie im Vergleich zu bestehenden Standards

Verknüpfung von Personen, Orten und Ereignissen

Implementierung der Datenbank

Datenbankdesign

Herausforderungen bei der Datenaufnahme

Verknüpfung mit externen Quellen (z.B. Wikidata)

Analyse der Netzwerke

Soziale Netzwerke des Vereinslebens

Verbindungen zwischen Mitgliedern

Kooperationen mit anderen Vereinen

Politische Netzwerke und deren Veränderungen

Einfluss der NS-Diktatur auf die Netzwerke

Feldpostkarten als Quelle für militärische Netzwerke

Geografische Ausdehnung der Netzwerke

Einsatzorte der Chormitglieder während des Krieges

Lokale und überregionale Verbindungen

Diskussion der Ergebnisse

Sichtbarmachung der Netzwerke durch Nodegoat und Netzwerkanalyse

Gibt es Veränderungen der Netzwerke im historischen Kontext?

Bibliographie

Referenzen, die noch nachzuschauen sind:

– Akten_Gesamtübersicht.csv im Anhang

References

„78th Sturm-Division (Wehrmacht)“. Unter Mitarbeit von Sven Burkhardt, (Zugriff am besucht am 12. März 2025)

. Besucht am 12. März 2025. <https://www.wikidata.org/wiki/Q125489568>.

Altenburger, Andreas. „Lexikon der Wehrmacht“, (Zugriff am besucht am 15. Januar 2023)

. Besucht am 15. Januar 2023. <https://www.lexikon-der-wehrmacht.de/Gliederungen/Infanteriedivisionen/205ID.htm>.

Burkhardt, Sven. „ArcGIS StoryMaps“. ArcGIS StoryMaps, (Zugriff am besucht am 12. März 2025)

. Besucht am 12. März 2025. <https://storymaps.arcgis.com>.

„DRK Suchdienst | Suche per Feldpostnummer“. DRK Suchdienst; Suche per Feldpostnummer. Unter Mitarbeit von Christian Reuter, (Zugriff am besucht am 12. März 2025)

. Besucht am 12. März 2025. <https://vbl.drk-suchdienst.online/Feldpostnummer/FPN.aspx>.

„Forum Geschichte der Wehrmacht“. Forum Geschichte der Wehrmacht. Unter Mitarbeit von Dieter Hermans, (Zugriff am besucht am 12. März 2025)

. Forum. Besucht am 12. März 2025. <https://www.forum-der-wehrmacht.de/>.

Garoufallou, Emmanouel und María-Antonia Ovalle-Perandones, Hrsg. *Metadata and Semantic Research. 14th International Conference, MTSR 2020, Madrid, Spain, December 2–4, 2020. Revised Selected Papers*. Bd. 1355. Communications in Computer and Information Science. Madrid, Spain: Springer Nature Switzerland AG, 2. Dezember 2020

. ISBN: 978-3-030-71903-6, besucht am 5. Juli 2025. https://basel.swisscovery.org/discovery/openurl?institution=41SLSP_UBS&vid=41SLSP_UBS:live&doi=10.1007%2F978-3-030-71903-6_30.

Gemeinde Murg, Hrsg. *Geschichte Gemeinde Murg*

. Besucht am 29. Juni 2025. <https://www.murg.de/seite/33378/geschichte.html>.

Hartmann, Christian. *Wehrmacht im Ostkrieg - Front und militärisches Hinterland 1941/42*. 2. Auflage. Bd. 75. Quellen und Darstellungen zur Zeitgeschichte Herausgegeben vom Institut für Zeitgeschichte. München: R. Oldenbourg Verlag, 2010

.

Hollmann, Prof. Dr. Michael. „Freiburg“. Bundesarchiv Freiburg im Breisgau (Abteilung Militärarchiv), (Zugriff am besucht am 12. März 2025)

. Besucht am 12. März 2025. <https://www.bundesarchiv.de/das-bundesarchiv/standorte/freiburg/>.

„Lexikon der Wehrmacht“. Unter Mitarbeit von Andreas Altenburger, (Zugriff am besucht am 12. März 2025)

. Besucht am 12. März 2025. <http://www.lexikonderwehrmacht.de/>.

Martinez, Roxana und Gonzalo Pereyra Metnik. „Comparative Study of Tools for the Integration of Linked Open Data: Case study with Wikidata Proposal“.

„OWL Web Ontology Language Guide“. Unter Mitarbeit von Michael K. Smith, Chris Welty und Deborah L. McGuinness, (Zugriff am besucht am 5. Juli 2025)

. Besucht am 5. Juli 2025. <https://www.w3.org/TR/owl-guide/>.

Rass, Christoph und René Rohrkamp. *Deutsche Soldaten 1939-1945 Handbuch einer biographischen Datenbank zu Mannschaften und Unteroffizieren von Heer, Luftwaffe und Waffen-SS*. Aachen, 2009

.

Tessin, Georg. *Verbände und Truppen der deutschen Wehrmacht und Waffen-SS im Zweiten Weltkrieg 1939-1945*. Bd. Band 1 - Die Waffengattungen — Gesamtübersicht. Osnabrück: HIBLIO Verlag, 1977

.

„WGS84 | Landesamt für Geoinformation und Landesvermessung Niedersachsen“. Landesamt für Geoinformation und Landesvermessung Niedersachsen, (Zugriff am besucht am 5. Juli 2025)

. Besucht am 5. Juli 2025. https://www.lgln.niedersachsen.de/startseite/wir_uber_uns/hilfe_support/lgln_lexikon/w/wgs84-190576.html.