



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



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, die beteiligten Personen sowie deren Netzwerke und dessen geographische Ausdehnung sichtbar zu machen.

Inhaltsverzeichnis

Einleitung	2
Ziel und Relevanz der Arbeit	2
Formulierung der Forschungsfrage	2
Aufbau der Arbeit	2
Geografischer und historischer Kontext	2
Korpus	3
Quellen	3
Quellentradierung	3
Quellenbeschrieb	4
Datierung der Quelle	4
Dokumententyp	4
Sichtung & Kategorisierung in Akten	4
Digitalisierung der Quellen	4
Transkription	5
Tagging	6
Tagging mit Transkribus	6
Tagging mit LLM	6
Export	6
Forschungsstand und Forschungslücke	6
Methodisches Vorgehen	9
Genutze Tools	9
LOD – Linked Open Data	9
Protégé	11
GraphDB	11
mma-Ontologie	12
Wikidata	13
GeoNames	14
Transkriptionen (Methodenvergleich)	14
Tesseract	14
LLM	14
Transkribus	15

Large Language Models	17
Msty	17
Alphabet – Gemini	18
Anthopic – Claude	18
OpenAI – ChatGPT	18
Nodegoat	18
Netzwerkanalyse als Methode	19
Theoretischer Hintergrund der Netzwerkanalyse	19
Ziele der Netzwerkanalyse im Kontext der Quellen	19
Technische Umsetzung (Tools, Datenbankstruktur)	19
Pipeline	19
Aufbau XML to JSON Pipeline	19
Übersichtsgrafik der Pipeline	19
Module im Detail	19
document_schemas.py	19
__init__.py	19
Person-Matcher	21
place_matcher.py	22
organization_matcher.py	22
letter_metadata_matcher.py	22
type_matcher.py	22
event_matcher.py	22
date_matcher.py	22
Assigned_Roles_Module.py	22
unmatched_logger.py	22
KEINE AHNUNG WAS DIE HIER MACHEN	23
validation_module.py	23
validation_module.py	23
test_role_schema.py	23
llm_enricher.py	23
enrich_pipeline.py	23
Analyse & Diskussion der Ergebnisse	23

Visualisierung auf der VM	23
Fazit und Ausblick	23
Zusammenfassung der zentralen Erkenntnisse	23
Methodische Herausforderungen und Lösungen	23
Ausblick auf zukünftige Forschung und mögliche Erweiterungen der Datenbank .	23
ALTER SCHEISS	24
Forschungsstand zu den Quellen	24
Beschreibung des Archivbestands	24
Methodischer Zugang	25
Digitale Erfassung und Strukturierung der Quellen	25
Gliederung in Akten	25
Digitalisierung und Transkription	25
Tagging in Transkribus	25
Digitalisierungsprozess und Herausforderungen	26
Diagram Pipelineübersicht	31
Gründe für den Wechsel zu Nodegoat	33
Nodegoat Modellierung	33
Netzwerkanalyse als Methode	33
Theoretischer Hintergrund der Netzwerkanalyse	33
Ziele der Netzwerkanalyse im Kontext der Quellen	33
Technische Umsetzung (Tools, Datenbankstruktur)	33
Aufbau der Datenbank	34
Konzeption der Datenmodellierung	34
Eigene Ontologie im Vergleich zu bestehenden Standards	34
Verknüpfung von Personen, Orten und Ereignissen	34
Implementierung der Datenbank	34
Datenbankdesign	34
Herausforderungen bei der Datenaufnahme	34
Verknüpfung mit externen Quellen (z.B. Wikidata)	34
Analyse der Netzwerke	35
Soziale Netzwerke des Vereinslebens	35

Verbindungen zwischen Mitgliedern	35
Kooperationen mit anderen Vereinen	35
Politische Netzwerke und deren Veränderungen	35
Einfluss der NS-Diktatur auf die Netzwerke	35
Feldpostkarten als Quelle für militärische Netzwerke	35
Geografische Ausdehnung der Netzwerke	35
Einsatzorte der Chormitglieder während des Krieges	35
Lokale und überregionale Verbindungen	35
Diskussion der Ergebnisse	36
Sichtbarmachung der Netzwerke durch Nodegoat und Netzwerkanalyse	36
Gibt es Veränderungen der Netzwerke im historischen Kontext?	36
Bibliographie	37
References	37
Anhang	40
PDF_to_JPEG.py	40
Tagging in Transkribus	41
Inhaltliche Tags	42
Strukturelle Tags	44

Einleitung

Ziel und Relevanz der Arbeit

Formulierung der Forschungsfrage

Aufbau der Arbeit

Geografischer und historischer Kontext

Die vorliegende Arbeit stützt sich auf Unterlagen aus dem Archiv des „Männerchor Murg“ dessen Nachfolge im Jahr 2021 durch die „New Gospelsingers Murg“ angetreten wurde. Murg ist eine deutsche 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, während der Rhein mit seiner Drahtseil-Fähre eine bedeutende Verkehrs- und Handelsverbindung bot, die bis zum Ersten Weltkrieg privat betrieben wurde.

Mit dem Ausbau der Landstrasse, der heutigen Bundesstrasse 34, sowie dem Anschluss an die Bahnstrecke Basel–Konstanz entwickelte sich Murg im 19. Jahrhundert von einer landwirtschaftlich geprägten Siedlung zu einer Gewerbe-, Handels- und Industriegemeinde. Die Wasserkraft wurde dabei zu einem entscheidenden Standortfaktor: Die Ansiedlung der Schweizer Textilfirma *Hüssy & Künzli AG* im Jahr 1853¹ trug wesentlich zum wirtschaftlichen Wachstum der Gemeinde bei. Zahlreiche Arbeitskräfte, vor allem aus der benachbarten Schweiz, machten Murg zu einem wichtigen Standort der regionalen Textilindustrie.

Die Gründung des *Männerchor Murg* im Jahr 1861 durch Schweizer Textilarbeiter belegt diesen engen Zusammenhang zwischen wirtschaftlicher Migration, Industrialisierung und lokalem Vereinswesen. Diese historische Verflechtung bildet eine zentrale Grundlage für die vorliegende Untersuchung.

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

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 1945, dem Tag der bedingungslosen Kapitulation Deutschlands.

Die zeitliche Eingrenzung ist notwendig, um die Funktionalität der im Folgenden beschriebenen Pipeline in einem klar definierten historischen Kontext demonstrieren zu können. Gleichzeitig führt diese Auswahl zu einer bewussten Reduzierung der potenziell erfassten Akteurinnen und Akteure, Orte und Organisationen. Diese Fokussierung ist insbesondere im Hinblick auf die Erstellung einer verlässlichen Groundtruth bedeutsam, die durch ergänzende Archivrecherchen mit historischen Metadaten angereichert wird.

Die Kombination aus einer präzise definierten Quellengrundlage und der digitalen Anreicherung dient dazu, das Potenzial der computergestützten Auswertung historischer Dokumente exemplarisch aufzuzeigen. Zugleich unterstreicht sie, dass die Qualität der Ergebnisse wesentlich von der sorgfältigen Eingrenzung des Korpus und der manuellen Validierung und Anreicherung abhängt.

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

Datierung der Quelle

blabla Hier eine Grafik über die Verteilung im Ordner einfügen.

Dokumententyp

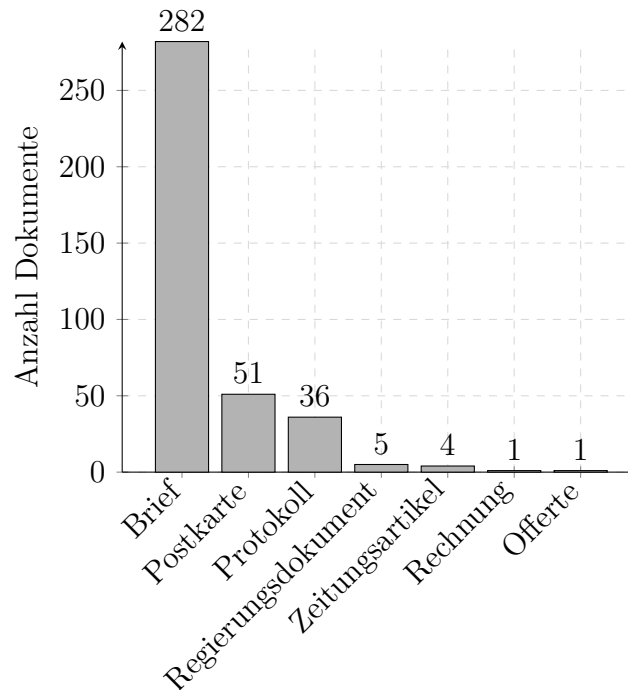


Abbildung 1: Verteilung der Dokumententypen im untersuchten Bestand (150 Akten – 381 Seiten).

Sichtung & Kategorisierung in Akten

blabla

Digitalisierung der Quellen

Die vorhandenen analogen Dokumente müssen zunächst fachgerecht für den Digitalisierungsprozess aufbereitet werden. Hierzu werden die Akten aus ihren ursprünglichen Ablesesystemen entnommen und sorgfältig von Heftklammern, Büro- und Gummibändern befreit. Diese konservatorischen Maßnahmen sind notwendig, um die langfristige Materialerhaltung zu gewährleisten, da insbesondere Korrosionsspuren ehemaliger Metallklammern die Papierfasern nachhaltig schädigen können. Zudem finden sich häufig Anzeichen von Säurefraß, sofern nicht säurefreies Archivmaterial verwendet wurde.

Für die eigentliche Digitalisierung kommt die native „Dateien“-Applikation von Apple² zum Einsatz. Diese bietet neben einer vergleichsweise hochauflösenden Erfassung die Möglichkeit zur direkten Speicherung in einem Cloud-basierten Speichersystem sowie eine automatische Texterkennung (OCR). Ziel dieser Vorgehensweise ist es, die digitalisierten Inhalte möglichst schnell durchsuchbar zu machen und standortunabhängig für das Projekt zugänglich zu machen.

Die Aufnahme der Dokumente erfolgt mithilfe eines Tablets, das auf einem stabilen Stativ exakt im rechten Winkel (90°) über dem zu digitalisierenden Schriftgut positioniert wird. Diese einfache, jedoch effiziente Konfiguration gewährleistet eine gleichbleibend hohe Bildqualität bei gleichzeitig hoher Verarbeitungsgeschwindigkeit. Die digital erfassten Dateien werden konsistent benannt und folgen einer vorab definierten Gesamtübersicht der Bestände. Mehrseitige Konvolute werden dabei als zusammengehörige Akteneinheiten geführt, während Einzeldokumente entsprechend separat erfasst werden. Die Archivierung erfolgt sowohl analog als auch digital auf Seitenebene, um eine möglichst feingranulare Erschließung zu ermöglichen.

Die initiale Speicherung erfolgt dabei standardmäßig im PDF-Format. Für die anschließende Verarbeitung mit den unten dargestellten Transkriptionswerkzeugen müssen die Dokumente jedoch in das JPEG-Format konvertiert werden. Die Umwandlung erfolgt automatisiert mithilfe eines eigens erstellten Python-Skripts, wie in Anhang beschrieben.³ Es extrahiert die Seiten, speichert im geeigneten Format ab und ergänzt die Dateinamen systematisch um eine dreistellige, führend nullengefüllte Seitennummer.

Transkription

Für die Transkription der Daten wurde ein best-practise Ansatz gewählt. Nach Tests mit dem Python-Modul „*Tesseract*“ und unterschiedlichen LLMs wurde auf Transkribus zurückgegriffen. Eine Gegenüberstellung der drei erwähnten Tools findet sich im Kapitel [Transkriptionen \(Methodenvergleich\)](#)

2. Vgl. [Apple Support: Dateien-App](#)

3. Sven Burkhardt, *github/PDF_to_JPEG.py*, Version 1.0, Computer software, Basel, 23. April 2025, besucht am 23. April 2025, https://github.com/Sveburk/masterarbeit/blob/main/3_MA_Project/Hilfs_Scripte/JPEG_to_PDF.py.

Tagging

blabla

Tagging mit Transkribus

blabla

Tagging mit LLM

blabla

Export

blabla

Forschungsstand und Forschungslücke

Die vorliegende Arbeit knüpft an mehrere Vorarbeiten an, die in den Jahren 2018 und 2022 am Departement Geschichte der Universität Basel durchgeführt werden. In zwei Transkribus-Seminaren werden erste Teilbestände der *Männerchor Akten 1925–1944* erschlossen und in einem Korpus von 137 Einzeldokumenten zusammengeführt⁴. Sie werden mit Metadaten versehen: Seitenlage im Ordner, Kurztitel und Entstehungsdatum, um die Grundlage für eine systematische Erschließung zu schaffen.

Während in einem frühen Projektschritt vorrangig häufig genannte Personennamen („Carl Burger“, „Fritz Jung“) dokumentiert werden, richtet sich der Fokus im zweiten Schritt auf die Feldpost. Ziel ist es, über die Auswertung der Feldpostnummern Rückschlüsse auf beteiligte Militäreinheiten, deren Stationierungen und Verlagerungen während des Zweiten Weltkriegs zu ziehen.

Für diese Recherchen kommen einschlägige Fachliteratur zu den jeweiligen Fachgebieten zum Einsatz. Hier sind besonders die Bücher von Alex Buchner⁵, Christian Hartmann⁶,

4. Vgl. Sven Burkhardt, „Feldpost an den Männerchor Murg - Storymaps“, ArcGIS StoryMaps, (Zugriff am besucht am 12. März 2025)

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

5. Vgl. Alex Buchner, *Das Handbuch der Deutschen Infanterie 1939 – 1945*, 2. Aufl. (Friedberg: Podzun-Pallas, 1989

), ISBN: 3-7909-0301-9.

6. Vgl. 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

Werner Haupt⁷, Christoph Rass⁸, Georg Tessin⁹ und Christian Zentner¹⁰ zu nennen.

Darüberhinaus werden eigene Recherchen in den Beständen des *Bundesarchivs – Militärarchiv Freiburg*¹¹ durchgeführt. Ergänzende Recherchen stammen aus den Suchlisten des *Deutschen Roten Kreuzes (DRK)*¹². Hinzu kommen philatelistische Übersichts-Websites¹³, die bei der Entzifferung von Briefmarken und Stempeln helfen. Absolut essentiell für den Erfolg dieser Recherchen sind Citizen-Science-Foren¹⁴. Sie ergänzen und validieren eigene Forschung.

Parallel zur inhaltlichen Erschließungen entsteht 2022 eine erste digitale Storymap mit *ArcGIS*, die zentrale Ergebnisse des Projekts öffentlich zugänglich macht. Grundlage bildet die Sichtung, konservatorische Aufbereitung und Digitalisierung von zunächst rund 30 der etwa 800 Seiten Vereinsakten. Der Teilkorpus wird entheftet, gescannt und mit Metadaten wie Absender, Datum, Feldpostnummer und Einheit versehen. Da jedes Dokument einen anderen Verfasser aufweist, erfolgt die Transkription manuell. Eine automatische Handschriftenerkennung ist aufgrund der heterogenen Schriftbilder nicht praktikabel. Am Beispiel einzelner Säger wie *Emil Durst* lässt sich durch die Rechercheergebnisse mithilfe der Feldpostnummern und ergänzender Kartenmaterialien der Aufenthaltsort bis auf Gebäude oder wenige Meter genau rekonstruieren. Diese Erkenntnisse werden mit

).

7. Vgl. Werner Haupt, *Das Buch der Infanterie*, 1. Aufl. (Friedberg, Hanau: Podzun-Pallas, 1982), ISBN: 3-7909-0176-8.

8. Vgl. 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)

).

9. 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)

).

10. Vgl. Christian Zentner, *Illustrierte Geschichte des Zweiten Weltkriegs* (München: Südwest Verlag GmbH, 1983)

).

11. 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/>.

12. „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>.

13. „Feldpost Number Database | GermanStamps.net“, (Zugriff am besucht am 9. Juli 2025)

, besucht am 9. Juli 2025, <https://www.germanstamps.net/feldpost-number-database/>.

14. vor Allem werden verwendet: *Forum der Wehrmacht* („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/>) und das *Lexikon der Wehrmacht* (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>).

historischen Karten, Luftbildern und Ortsrecherchen verknüpft und in einer interaktiven ArcGIS-Karte visualisiert, die Stationierungen, Märsche und Frontverschiebungen der Chormitglieder anschaulich darstellt.

Die in diesen Vorprojekten erarbeiteten Listen, Geodaten, Transkriptionen und Visualisierungen fließen in die vorliegende Arbeit ein und bilden eine wesentliche Grundlage für die erweiterte, automatisierte Pipeline, die im Folgenden vorgestellt wird. Dazu gehören beispielsweise auch die Verbandsabzeichen, Taktische Zeichen¹⁵ der jeweiligen Einheiten, die auch in die Groundtruth der vorliegenden Arbeit inkorporiert werden.

Abgesehen von diesen Vorarbeiten ist der Quellenkorpus wissenschaftlich unerschlossen. Mit dieser Arbeit liegt erstmals eine umfassendere wissenschaftliche Auswertung vor.

Mit der notwendigen manuellen Recherche in oben dargelegten Datenbankstrukturen wird zugleich sichtbar, wie sehr es an Brücken fehlt, um unterschiedliche Klassifikationen, fachspezifische Ordnungslogiken und semantische Webtechnologien nachhaltig miteinander zu verbinden. Ein verhältnismässig einfaches Webscraping nach Informationen zu diesem Korpus ist nahezu unmöglich. Ausgeführt werden diese Probleme beispielsweise bei Smiraglia und Scharnhorst (2021)¹⁶, die anhand konkreter Fallstudien verdeutlichen, wie fragmentiert semantische Strukturen bislang entwickelt werden und welche Hürden bei der praktischen Verknüpfung heterogener Wissensorganisationen bestehen. Dabei benennen sie insbesondere die Herausforderungen bei der Übersetzung historisch gewachsener Klassifikationen in standardisierte semantische Formate, die Notwendigkeit dauerhafter technischer Wartung und die Abhängigkeit von nachhaltigen Infrastruktur-Partnern¹⁷.

Für eine Einordnung zu historischen Netzwerkanaylsen sei auf Gamper&Reschke¹⁸ verwiesen. Der Sammelband *Knoten und Kanten III* verdeutlicht, dass die historische Netzwerkanalyse zwar von einem interdisziplinär etablierten Methodenkanon profitiert, jedoch nach wie vor erheblichen Herausforderungen steht. Dazu zählen die Fragmentierung

15. Vgl Haupt, *Das Buch der Infanterie*, S.64-66.

16. Vgl. Smiraglia Richard und Scharnhorst Andrea, *Linking Knowledge. Linked Open Data for Knowledge Organization and Visualization*, Version Number: editorsversion, prior to publication (Zenodo, 3. Mai 2022

), besucht am 14. Januar 2025, <https://doi.org/10.5771/9783956506611>, <https://zenodo.org/records/6513663>.

17. Vgl. Richard und Andrea, Kap. 2 und 5.

18. Markus Gamper und Linda Reschke, *Knoten und Kanten III: Soziale Netzwerkanalyse in Geschichts- und Politikforschung*, hrsg. Martin Düring (transcript Verlag, 27. April 2015

), ISBN: 978-3-8394-2742-2, besucht am 14. Januar 2025, <https://doi.org/10.1515/9783839427422>, <https://www.degruyter.com/document/doi/10.1515/9783839427422/html>.

historischer Quellen, der hohe manuelle Erfassungsaufwand und methodische Desiderate im Umgang mit zeitlichen und räumlichen Dimensionen. Erschwerende Faktoren einer systematischen Erfassung relationaler Strukturen. Dennoch eröffnen netzwerkanalytische Verfahren – besonders im Zusammenspiel mit relationaler Soziologie und Figurationsansätzen – neue Perspektiven auf Macht, Abhängigkeiten und Akteurskonstellationen in historischen Gesellschaften.

Methodisches Vorgehen

Genutzte 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 verbinden¹⁹. Wesentliches

19. 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)

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.²⁰ 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 unterschiedlichen Quellen unter verschiedenen URIs²¹ geführt werden, als dasselbe reale Objekt²² und ermöglicht so eine präzise Zusammenführung von Informationen — ein Grundpfeiler für die Interoperabilität im Semantic Web. Die OWL-Spezifikation baut auf RDF²³ auf und erweitert es um zusätzliche Konzepte. Die RDF-Daten werden häufig im Turtle-Format (TTL) serialisiert, einer textbasierten Notation für RDF, die eine kompakte, leicht lesbare Schreibweise bietet. Dieses Format eignet sich besonders für den Austausch und die manuelle Bearbeitung von RDF-Tripeln. Die Sprache liegt in drei Varianten vor²⁴, die sich im Grad ihrer Ausdrucksstärke unterscheiden.²⁵ 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.²⁶ Für die technische Umsetzung für diese Arbeit werden zwei zentrale Werkzeuge genutzt: Protégé zur Modellierung der Ontologie und GraphDB für deren Verwaltung und Abfrage.

), 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.

20. 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/>.

21. Abk. URI Uniform Resource Identifier

22. vgl. „OWL Guide“, 2.3. Data Aggregation and Privacy.

23. Abk. RDF Resource Description Framework

24. OWL Lite, OWL DL und OWL Full

25. vgl. „OWL Guide“, 1.1. The Species of OWL.

26. vgl. Garoufallou und Ovalle-Perandones, *Metadata and Semantic Research*, S. 14.

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 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 erfasst²⁷. Beim Schreiben dieser initialen Ontologie aus rund 2000 Zeilen Code erweist sich schnell ein neues Problem. Die Datengrundlage aus den geschilderten Vorprojekten (siehe [Forschungsstand und Forschungslücke](#)) ist zu klein, um daraus eine aussagekräftige Netzwerkanalyse zu machen. Hierfür erweisen sich die Unterschiede der Daten zusätzlich als zu grosse Grundlage des Global und damit aufwendig. Der Fokus der Arbeit verschiebt sich dementsprechend von der Ontologieentwicklung auf die Extraktion von Entitäten.

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“²⁸ und der „Suchdienst des DRK“²⁹. In beiden Fällen liegen die Daten jedoch nicht als LOD vor, sondern im Forum als einfache Strings und beim Deutschen Roten Kreuz als OCR-PDF³⁰ 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 aufgegriffen³¹.



Abbildung 2: Ausschnitt der TTL-Ontologie.

27. Abk. mmma; Männerchor Murg Masterarbeit.

28. vgl. Altenburger, „[Lexikon der Wehrmacht](#)“.

29. vgl. „[DRK Suchdienst | Suche per Feldpostnummer](#)“.

30. OCR = Optical Character Recognition

31. siehe Kapitel Nodegoat

Wikidata

Wikidata³² 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.“*³³

Wikidata entspricht, ebenso wie das nachfolgend beschriebene GeoNames, den FAIR-Prinzipien: Die Daten sind **F**indable und **A**ccessible, **I**nteroperable und **R**eusable³⁴.

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. Für diese Arbeit eigens angelegter Einträge auf Wikidata werden trotz systematischer Verknüpfung mit anderen dort verwalteten Entitäten, etwa mit Armeen, Militäreinheiten, Orten und Personen, entfernt die Community-Moderation etwa 70% dieser Einträge. Das zeigt einerseits hohe internen Qualitätsanforderungen auf, andererseits werden diese jedoch nicht klar kommuniziert. Mit regidem Löschen neuer Einträge wird die Verlässlichkeit und den Nutzen der geleisteten Arbeit erheblich begrenzt. Aufwand und Unsicherheit über die Persistenz der Einträge machen den ursprünglich vorgesehenen LOD-Ansatz in dieser Form nicht praktikabel.

32. Vgl. „Wikidata“, (Zugriff am besucht am 5. Juli 2025)

, besucht am 5. Juli 2025, https://www.wikidata.org/wiki/Wikidata:Main_Page.

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

34. vgl. Mark D. Wilkinson u. a., „The FAIR Guiding Principles for scientific data management and stewardship“, Publisher: Nature Publishing Group, *Scientific Data* 3, Nr. 1 (15. März 2016

): S. 2, ISSN: 2052-4463, besucht am 5. Juli 2025, <https://doi.org/10.1038/sdata.2016.18>, <https://www.nature.com/articles/sdata201618>.

GeoNames

Ebenso wie Wikidata bietet *GeoNames*³⁵ 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-System³⁶ 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.

GeoNames wird in dieser Arbeit intensiv zur Referenzierung von Ortsnamen verwendet und bildet die Basis für die Groundtruth, wie sie in den Kapiteln *Nodegoat* und *place_matcher.py* beschrieben ist. Im Gegensatz zu Wikidata wurde hier von Beginn an darauf verzichtet, eigene Ortsdatensätze zu ergänzen. Dies liegt einerseits an den klar kommunizierten Community-Guidelines und andererseits daran, dass der Datensatz bis auf wenige, sehr lokale Flurnamen als nahezu vollständig gilt.

Historische Gebäude wie Gaststätten oder Spitäler fehlen folgerichtig in der GeoNames-Datenbank. Diese Lücke ist erwartbar, aber erwähnenswert, da GeoNames ansonsten eine nahezu vollständige und ausgesprochen detaillierte Datengrundlage bietet.

Transkriptionen (Methodenvergleich)

Tesseract

war scheisse

LLM

war scheisse und manipulativ

35. **noauthor_geonames_nodate.**

36. *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.

Transkribus

Transkribus ist eine webbasierte Plattform zur automatisierten Handschrifterkennung (HTR) und Texterkennung (OCR), die sich seit ihrer Entwicklung im EU-Projekt READ (Recognition and Enrichment of Archival Documents)³⁷ als Standardwerkzeug in den digitalen Geschichtswissenschaften etabliert hat³⁸. Betrieben wird Transkribus durch die READ-COOP SCE, einer europäischen Genossenschaft.

Für die vorliegende Arbeit wird sowohl die Webanwendung „*Transkribus Lite*“ als auch der *Desktop-„Transkribus Expert Client“* genutzt. Letzterer ist eine Fortführung der ursprünglichen, im Rahmen des READ-Projekts entwickelten Software. Die Verwendung des Expert Clients in Kombination mit *FileZilla Pro* als FTP-Client ermöglicht es, die grosse Zahl an Scanseiten effizienter in das Transkribus-Dateimanagement einzupflegen. Die für das Projekt nötige Scholar-Lizenz wird über RISE der Universität Basel³⁹ bezogen.

Neben der eigentlichen Transkription bietet Transkribus die Möglichkeit, Named Entities wie Personen, Orte, Organisationen und Datumsangaben direkt im Dokument zu annotieren. Dies geschieht über eine umfangreiche Tagging-Funktion, die neben struktureller Tags wie Abkürzungen („abbrev“) auch individuell erweiterbar sind. Für diese Arbeit wird beispielsweise der Custom-Tag **signature** eingeführt, der dem Anhang entnommen werden können.

Nach erfolgter Bearbeitung stellt Transkribus Funktionen zur Verfügung, die Transkriptionen inklusive Tags als strukturiertes Markup im XML-Format zu exportieren.

In der praktischen Anwendung zeigt sich jedoch eine erhebliche Diskrepanz zwischen den vorgesehenen Funktionen und der tatsächlich umgesetzten Exportlogik. Die händisch markierten und ausformulierten Abkürzungen werden im generierten Page-XML nicht ausgegeben, sodass diese Information für eine weitergehende Auswertung verloren geht.

Aus Abkürzungen wie V.D.A soll in XML `<abbrev expansion=„Verein für das Deutschtum im Ausland`

37. Vgl. „Recognition and Enrichment of Archival Documents | READ | Projekt | Fact Sheet | H2020“, CORDIS | European Commission, (Zugriff am besucht am 6. Juli 2025), besucht am 6. Juli 2025, <https://cordis.europa.eu/project/id/674943>.

38. Vgl. Günter Mühlberger, „Transkribus Eine Forschungsplattform für die automatisierte Digitalisierung, Erkennung und Suche in historischen Dokumenten“ (Kolloquium der ETH-Bibliothek, Zürich, 25. April 2019

), postnote, besucht am 6. Juli 2025, https://ethz.ch/content/dam/ethz/associates/ethlibrary-dam/documents/Aktuell/Veranstaltungen/17-15-Kolloquium/2019-04-29_17-15-Kolloquium_transkribus.pdf.

39. Weitere Informationen: Eric Decker, „Home | RISE | Research & Infrastructure Support | Universität Basel“, Research & Infrastructure Support, (Zugriff am besucht am 6. Juli 2025), postnote, besucht am 6. Juli 2025, <https://rise.unibas.ch/de/>.

werden. Tatsächlich fehlt der Tag komplett, V.D.A wird zum einfachen String. Unter Umständen wird dieser nun von den LLMs nicht mehr als Organisation erkannt.

Besonders gravierend ist das Defizit bei der Handhabung von Listen. Zwar können Listenobjekte innerhalb der Benutzeroberfläche manuell angelegt und befüllt werden, beim Export bleibt die zugehörige XML-Struktur jedoch leer. Ein technischer oder benutzerfreundlicher Workaround, um diese Daten maschinenlesbar zu extrahieren, existiert bislang nicht. Um auf die Listeninhalte im XML nicht zu verzichten, müssen sie daher als regulärer Fliesstext markiert werden — mit dem Effekt, dass die logische Gliederung, die etwa für Mitgliederverzeichnisse, Inventarlisten oder Aufstellungen forschungsrelevant ist, systematisch verloren geht. Dies führt zu erheblichem Mehraufwand und stellt eine signifikante methodische Limitation dar, insbesondere für Vorhaben, die auf einer automatisierten Weiterverarbeitung konsistenter Strukturdaten beruhen. Hinzu kommt der grosse Zeitaufwand für die manuelle Listengestaltung, die letztlich nicht zielführend verwertet werden kann.

Die Herausforderungen decken sich mit den Beobachtungen von Capurro et al.⁴⁰, die in ihrem Experiment mit einem mehrsprachigen und multi-autoren Handschriftenkorpus ebenfalls auf erhebliche Grenzen von Transkribus hinweisen: Sowohl die automatische Layout-Erkennung als auch die Handhabung komplexer Dokumentstrukturen erwiesen sich als fehleranfällig und machten umfangreiche manuelle Nachbearbeitung unvermeidbar. Auch der Versuch, HTR-Ergebnisse⁴¹ nachträglich automatisch zu verbessern (Postcorrection), brachte keine Verbesserung der Fehlerquote. Damit bestätigt sich der Befund, dass ein projektspezifisch trainiertes Modell – wie es hier mit vier Trainings-Iterationen umgesetzt wurde – zwar die CER senken kann, strukturelle Probleme jedoch ungelöst bleiben.

Der methodische Mehrwert von Transkribus liegt trotz dieser Limitierungen in der Möglichkeit, ein eigenes, auf den spezifischen Handschriftencorpus angepasstes HTR-Modell zu trainieren. Ein generisches Modell (*The German Giant I*) erzielt zu Beginn eine Character Error Rate (*CER*) von 8,3%. Durch das Training auf einer eigens erstellten Groundtruth mit 149 Seiten lässt sich ein projektspezifisches Modell entwickeln, das die CER auf 6,58%

40. Carlotta Capurro, Vera Provatorova und Evangelos Kanoulas, „Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Author Manuscript Collection“, *Heritage* 6, Nr. 12 (29. November 2023): 7482–7494, ISSN: 2571-9408, besucht am 6. Juli 2025, <https://doi.org/10.3390/heritage6120392>, <https://www.mdpi.com/2571-9408/6/12/392>.

41. Abk.: **HTR**, Handwritten Text Recognition

senkt. Auch für das Tagging selbst wurde eine umfassende Groundtruth für etwaige spätere Anwendungen (und als Vergleich für die LLMs) erstellt. Sie besteht aus insgesamt 100 Seiten händischer Annotation. Diese umfassende manuelle Nachbearbeitung stellt ein möglichst konsistentes, maschinenlesbares XML sicher. Aus diesem Grund werden auch Taggingregeln festgehalten, die später auch an das LLM weitergegeben werden.

Insgesamt zeigt sich, dass Transkribus als Plattform eine sehr hilfreiche Basis für grossvolumige Transkriptionsprojekte bietet. Die automatisierte Erkennung spart erhebliche Zeit, ersetzt jedoch weder editorische Sorgfalt noch eine manuelle Qualitätssicherung. Gerade für Forschungsprojekte, die auf präzisen Strukturdaten beruhen, wie hier für Netzwerkanalysen mit Nodegoat, bleibt eine kritische Reflexion der Tool-Limitierungen unerlässlich. Transkribus unterstützt ausgesprochen bei der initialen Transkription, ermöglicht das verständlichere Lesen einzelner Passagen – darauf aufbauend ergibt sich ein Kontext, der eine vollständige Transkription mit viel manueller Arbeit möglich macht. Für Netzwerkanalysen relevante Informationen wie Signaturen werden oft falsch transkribiert – und eine Zuordnung per Tag ist daher durch den Menschen effizienter, als für sehr kleine Beispielmengen ein spezialisiertes Neuronales Netzwerk zu trainieren.

Large Language Models

Ein zentrales Werkzeug bei der Verarbeitung der historischen Quellen ist die weiter unten näher beschriebene Python-Pipeline, die auf der Verarbeitung von XML-Dateien basiert. Vorgreifend sei erwähnt, dass diese XML-Verarbeitung ein Large Language Model (LLM) zum Custom-Tagging nutzt. Nebst dem Tagging stellt das Programmieren dieser Pipeline eine der Kernherausforderungen dieses Forschungsprojekts dar. Für das Tagging und die Entwicklung der Pipeline werden verschiedene Large Language Models intensiv getestet und eingesetzt.

Msty Um ein dafür geeignetes LLM zu evaluieren, werden zu Beginn des Projektes beispielhafte Prompts erstellt und deren Ergebnisse systematisch verglichen. Um diesen Vergleich zu erleichtern, wird die Desktop-Anwendung Msty⁴² eingesetzt. Zu den zentralen Funktionen gehören parallele Chatinterfaces („Parallel Multiverse Chats“), eine flexible Verwaltung lokaler Wissensbestände („Knowledge Stacks“)⁴³, sowie eine vollständige

42. Vgl. „Msty - Using AI Models made Simple and Easy“, (Zugriff am besucht am 6. Juli 2025), besucht am 6. Juli 2025, <https://msty.app/>.

43. Vgl. „Msty - Using AI Models made Simple and Easy“.

Offline-Nutzung ohne externe Datenübertragung. Msty dient dazu, verschiedene Modelle zu testen, durch die Parallel Multiverse Chats Antworten zu vergleichen und Konversationen strukturiert zu verzweigen und auszuwerten.

Wichtig ist, dass dies kein klassisches Benchmarking auf Basis vergleichbarer Resultate ist. Es wird zu diesem frühen Projektzeitpunkt weder systematisch überprüft, welche Qualität der jeweilige Codeteil hat, noch wird gemessen, wie viel Prozent der Named Entities jeweils richtig erkannt werden. Der direkte Vergleich der getesteten LLMs liefert jedoch schnell ein klares Bild, welches Modell sich am besten eignet. Beprobt werden die Folgenden Anbieter und Modelle:

Alphabet – Gemini

Anthropic – Claude

OpenAI – ChatGPT

Nodegoat

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

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`

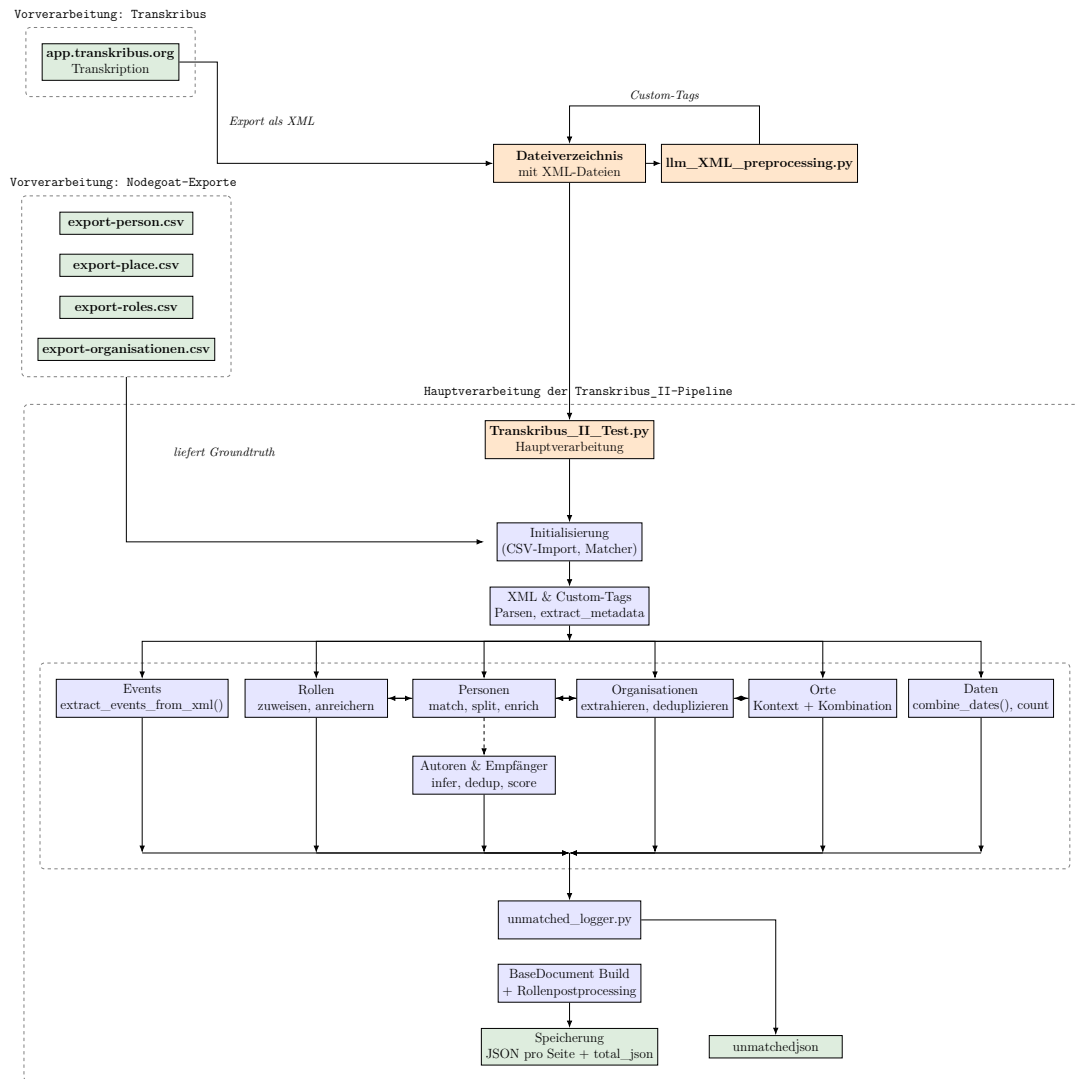


Abbildung 3: Übersicht der gesamten XML-to-JSON-Pipeline

Person-Matcher

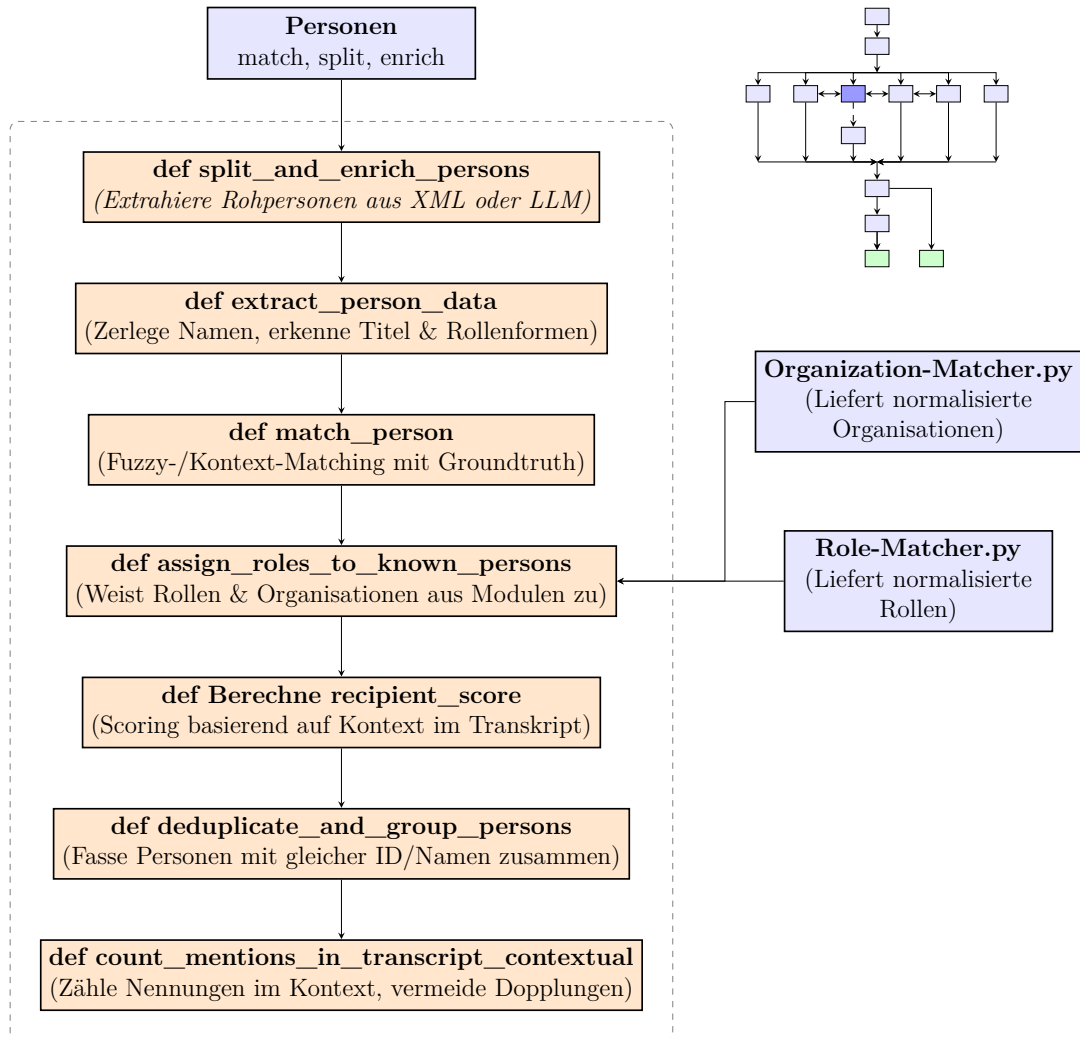


Abbildung 4: Detailliertes Prozessdiagramm für *Personen_matcher.py*

Oben rechts: Pipelineübersicht

Rechts: Input aus *role_matcher.py* und *organization_matcher.py*

`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`

Das Modul `unmatched_logger.py` dient der systematischen Protokollierung von Entitäten, die in der aktuellen Version der Groundtruth noch nicht enthalten sind. Diese Protokolle bilden die Grundlage für weiterführende Recherchen, durch die die Groundtruth schrittweise ergänzt und verbessert werden kann.

Innerhalb der Verarbeitungs-Pipeline wird das Modul `unmatched_logger.py` über die Funktion `process_single_xml()` im Hauptprogramm aufgerufen.

Bereits in der Testphase iteriert die Pipeline mehrfach über einzelne Segmente des Korpus, um den Code zu optimieren und alle bislang nicht erfassten Entitäten zuverlässig zu identifizieren.

Im Kern stellt das Modul die Funktion `log_unmatched_entities` bereit. Diese übernimmt die von den zuvor beschriebenen Matcher-Funktionen ermittelten Entitäten und prüft, ob sie in den entsprechenden Groundtruth-CSV-Dateien vorhanden sind.

Die Suche erfolgt iterativ innerhalb der Listenstrukturen für Personen, Orte, Rollen, Organisationen und Ereignisse. Wird eine Entität über ein XML-Custom-Tag einer dieser Kategorien zugewiesen, ohne dass sie in der Groundtruth verzeichnet ist, wird sie in einer spezifischen JSON-Datei protokolliert.

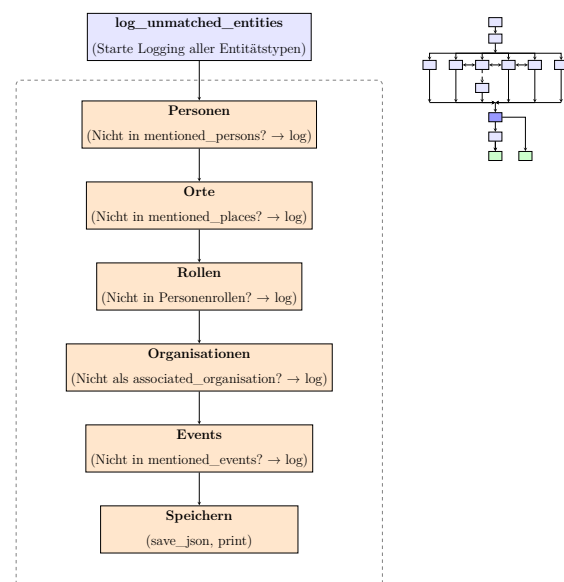


Abbildung 5:
Prozessdiagramm für
`unmatched_logger.py`
Oben rechts: Pipelineübersicht

Die folgenden Dateien werden dabei erzeugt:

- `unmatched__persons.json`
- `unmatched__places.json`
- `unmatched__roles.json`
- `unmatched__events.json`
- `unmatched__organisations.json`

Zusätzlich werden alle Einträge in einer zusammengeführten Datei `unmatched.json` gespeichert, um einen vollständigen Überblick über alle bislang nicht zugeordneten Entitäten zu gewährleisten.

Alle Ergebnisse werden zudem in einer Datei `unmatched.json` gespeichert, um einen Gesamtüberblick zu erhalten.

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 Akten⁴⁴. 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 vorliegende Arbeit stützt sich auf diese Vorarbeit und die darin gesammelten Daten. Beispielsweise werden die Feldpostbriefe um weitere Informationen ergänzt. Kernfragen hierfür sind: *Welche Einheiten verbergen sich hinter den Feldpostnummern? Wo waren die Einheiten, als der Brief geschrieben wurde?*

Hierzu werden Nachschlagetabellen in der Fachliteratur (vgl.⁴⁵), die Bestände des *Bundesarchives – Militärarchiv Freiburg*,⁴⁶ des *Suchdienstes des Deutschen Roten Kreuzes (DRK)*⁴⁷ sowie Citizen-Science-Projekte⁴⁸ herangezogen und letztere teils durch eigene Recherche ergänzt.

Für diese Arbeit wird die Kategorisierung von 2018 übernommen und auf den Seiten im Ordner erweitert.⁴⁹

Beschreibung des Archivbestands

44. Weiterführend vgl. Burkhardt, „Feldpost Storymaps“.

45. Tessin, *Verbände und Truppen*; Hartmann, *Wehrmacht im Ostkrieg*; Rass und Rohrkamp, *deutsche Soldaten 1939-1945*.

46. Hollmann, „Militärarchiv Freiburg“.

47. „DRK Suchdienst | Suche per Feldpostnummer“.

48. vgl. z.B. Wikidata: „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 2025)

, besucht am 12. März 2025, <http://www.lexikonderwehrmacht.de/>, „Forum Geschichte der Wehrmacht“.

49. Vgl. Sven Burkhardt, *github/Akten_Gesamtübersicht.csv*, 23. April 2025

, besucht am 6. Juli 2025, https://github.com/Sveburk/masterarbeit/blob/main/3_MA_Project/Da ta/Akten_Gesamt%C3%BCbersicht.csv.

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“-App⁵⁰ 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:

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

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 Archivkursen 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 Modell „The German Giant I“ mit einer CER von 8,30 % transkribiert. Mit insgesamt vier 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 genutzt 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 mit insgesamt 22.155 Wörtern Groundtruth, womit dann ein eigenes Transkribus-Modell⁵¹ (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:

51. „Transkribus“, Transkribus_Model_mmma, unter Mitarb. von Sven Burkhardt, (Zugriff am besucht am 25. Juni 2025)
, besucht am 25. Juni 2025, <https://app.transkribus.eu>.

Münch, 15. Aug. 41.

Mein lieber Hans!

Dein langer Brief ist mir sehr willkommen.
Ich würde mir ein Liedchen zu wünschen.
und kann mir die ganze Gelegenheit zuwenden.

Ich habe mir vorgenommen, heute um den
Mannchen Teestadt um den Titel des Liedchen
zu versuchen, so wie zum Beispiel um die
neuen „neue Liedchen“ überhaupt in der
musik beifügen, kein Aushaus. Vielleicht
gelingt es dir diesen Titel zu erhalten.

Dein Brief ist sehr schön und
„das alte Lied“ von dem. Ich will

Es wurde 1928 um 10. Teitub. Längere. Post
um Längere. Längere. in der Längere.
und wurde in der Längere. Längere.

Es ist sehr schön, das Liedchen zu finden.

Als Hans: „das alte Teestadt Liedchen“,
als Hans: „das alte Liedchen“ und vom
Liedchen in der Längere, das Hans Wahl.

Mit sehr. Längere

Ein

Carl.

Abbildung 6: 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 7: Transkription von [Abbildung 6](#)

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 8: Transkription durch ChatGPT von [Abbildung 7](#)

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:

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 Frankmarke 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 9: Tagging von [Abbildung 8](#)

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

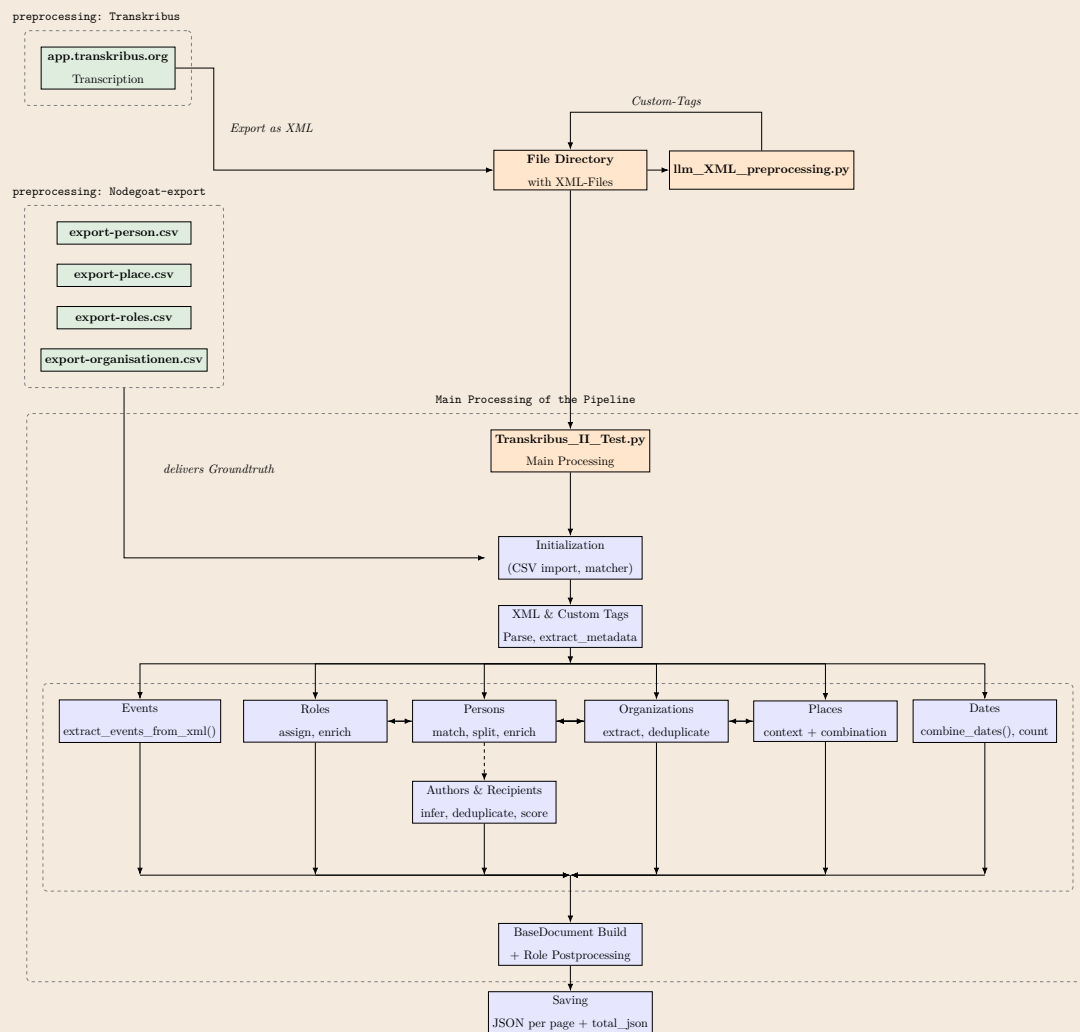
Vereinsführer des Männerchor

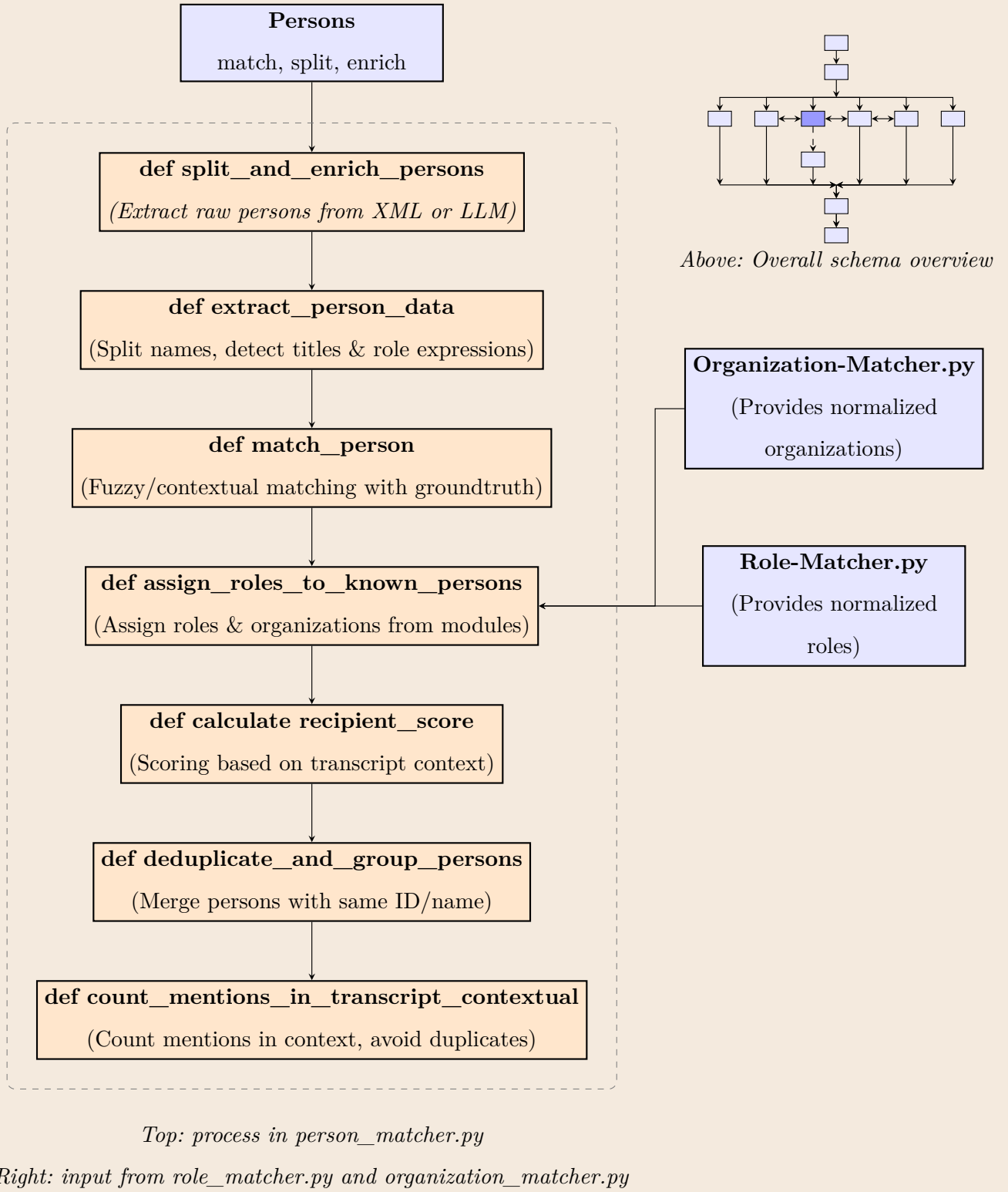
Murg

Laufenburg (Baden)

Rhina

Diagram Pipelineübersicht





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)

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

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>.
- Buchner, Alex. *Das Handbuch der Deutschen Infanterie 1939 – 1945*. 2. Aufl. Friedberg: Podzun-Pallas, 1989
. ISBN: 3-7909-0301-9.
- Burkhardt, Sven. „Feldpost an den Männerchor Murg - Storymaps“. ArcGIS StoryMaps, (Zugriff am besucht am 12. März 2025)
. Besucht am 12. März 2025. <https://storymaps.arcgis.com>.
- . *github/Akten_Gesamtübersicht.csv*, 23. April 2025
. Besucht am 6. Juli 2025. https://github.com/Sveburk/masterarbeit/blob/main/3_MA_Project/Data/Akten_Gesamt%C3%BCbersicht.csv.
- . *github/PDF_to_JPEG.py*. Version 1.0. Computer software. Basel, 23. April 2025
. Besucht am 23. April 2025. https://github.com/Sveburk/masterarbeit/blob/main/3_MA_Project/Hilfs_Scripte/JPEG_to_PDF.py.
- Capurro, Carlotta, Vera Provatorova und Evangelos Kanoulas. „Experimenting with Training a Neural Network in Transkribus to Recognise Text in a Multilingual and Multi-Author Manuscript Collection“. *Heritage* 6, Nr. 12 (29. November 2023): 7482–7494. ISSN: 2571-9408, besucht am 6. Juli 2025. <https://doi.org/10.3390/heritage6120392>. <https://www.mdpi.com/2571-9408/6/12/392>.
- Decker, Eric. „Home | RISE | Research & Infrastructure Support | Universität Basel“. Research & Infrastructure Support, (Zugriff am besucht am 6. Juli 2025)
. Besucht am 6. Juli 2025. <https://rise.unibas.ch/de/>.
- „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>.
- „Feldpost Number Database | GermanStamps.net“, (Zugriff am besucht am 9. Juli 2025)
. Besucht am 9. Juli 2025. <https://www.germanstamps.net/feldpost-number-database/>.
- „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/>.

- Gamper, Markus und Linda Reschke. *Knoten und Kanten III: Soziale Netzwerkanalyse in Geschichts- und Politikforschung*. Herausgegeben von Martin Düring. transcript Verlag, 27. April 2015
 . ISBN: 978-3-8394-2742-2, besucht am 14. Januar 2025. <https://doi.org/10.1515/9783839427422>. <https://www.degruyter.com/document/doi/10.1515/9783839427422/html>.
- 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
 .
- Haupt, Werner. *Das Buch der Infanterie*. 1. Aufl. Friedberg, Hanau: Podzun-Pallas, 1982
 . ISBN: 3-7909-0176-8.
- 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“.
- „Msty - Using AI Models made Simple and Easy“, (Zugriff am besucht am 6. Juli 2025)
 . Besucht am 6. Juli 2025. <https://msty.app/>.
- Mühlberger, Günter. „Transkribus Eine Forschungsplattform für die automatisierte Digitalisierung, Erkennung und Suche in historischen Dokumenten“. Kolloquium der ETH-Bibliothek, Zürich, 25. April 2019
 . Besucht am 6. Juli 2025. https://ethz.ch/content/dam/ethz/associates/ethlibrary-dam/documents/Aktuell/Veranstaltungen/17-15-Kolloquium/2019-04-29_17-15-Kolloquium_transkribus.pdf.
- „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
- .
- „Recognition and Enrichment of Archival Documents | READ | Projekt | Fact Sheet | H2020“. CORDIS | European Commission, (Zugriff am besucht am 6. Juli 2025)
. Besucht am 6. Juli 2025. <https://cordis.europa.eu/project/id/674943>.
- Richard, Smiraglia und Scharnhorst Andrea. *Linking Knowledge. Linked Open Data for Knowledge Organization and Visualization*. Version Number: editorsversion, prior to publication. Zenodo, 3. Mai 2022
. Besucht am 14. Januar 2025. <https://doi.org/10.5771/9783956506611>. <https://zenodo.org/records/6513663>.
- 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
- .
- „Transkribus“. Transkribus_Model_mmma. Unter Mitarbeit von Sven Burkhardt, (Zugriff am besucht am 25. Juni 2025)
. Besucht am 25. Juni 2025. <https://app.transkribus.eu>.
- „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.
- „Wikidata“, (Zugriff am besucht am 5. Juli 2025)
. Besucht am 5. Juli 2025. https://www.wikidata.org/wiki/Wikidata:Main_Page.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg u. a. „The FAIR Guiding Principles for scientific data management and stewardship“. Publisher: Nature Publishing Group, *Scientific Data* 3, Nr. 1 (15. März 2016
) : 160018. ISSN: 2052-4463, besucht am 5. Juli 2025. <https://doi.org/10.1038/sdata.2016.18>. <https://www.nature.com/articles/sdata201618>.
- Zentner, Christian. *Illustrierte GEschichte des Zweiten Weltkriegs*. München: Südwest Verlag GmbH, 1983
- .

Anhang

PDF_to_JPEG.py

```
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 QPixmap-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
34                    pix.save(output_path)
35                    # PDF-Datei schliessen
36                    doc.close()
37
38                # Erfolgsmeldung ausgeben
39                print(f"{file} wurde erfolgreich umgewandelt und gespeichert
40                ↪ in {dest_folder}")
41
42    # Pfade zu den Ordnern mit den PDF-Dateien (Quelle) und den JPG-Dateien (Ziel)
43    src_folder = r"/Users/svenburkhardt/Documents/D_Murger_Männer_Chor_Forschung/Scan_Mä_
44    ↪ nnerchor/Männerchor_Akten_1925-1945/Scan_Männerchor_PDF"
```

```

43 dest_folder = r"/Users/svenburkhardt/Documents/D_Murger_Männer_Chor_Forschung/Master_J
   ↳ arbeit/JPEG_Akten_Scans"
44
45
46 # Funktion aufrufen, um die Konvertierung durchzuführen
47 convert_pdf_to_jpg(src_folder, dest_folder)
48

```

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:

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 sind. 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>etwas unleserliches</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.