Aktueller Stand der Recherchen

Neu: Aufgesetzte Virtuelle Maschine mit IIIF Server zum Hosten der Daten unter https://dhlab-mmma.dhlab.unibas.ch/. mit Cantaloupe. Die Website wird diese Woche konfiguriert.

- Preprocessing von XML Dateien (Rohdaten von Tranksribus) nicht via OpenAI-API, sondern lokaler Lama LLM-Pipeline.

  —> zu zeitaufwendig

Recherche hat gerade keine Priorität, und läuft nebenbei… Keinerlei Literatur aktuell

immer noch korrekt

Darstellung von Zwischenergebnissen (sofern vorhanden)

Nodegoat:

- Erweiterung der Metadaten

Abgeschlossen

- Nodegoat-Datenbank dient durch CSV Export immer als Ground-Truth Grundlage für meine Analysen.

Coding:

Aufbau von Python-Pipeline mit mehreren Modulen, die die XML-Rohdaten aus Transkribus auswerten, gegen meine Groundtruth matchen (GT-CSV) und in JSON Format (für die Datenbank) umwandeln.

Modul	Kurzbeschrieb	Fehler/Fehlerquellen
Personen	Vornamen, Nachnamen, Spitznamen im Doc, gematched mit Ground Truth	Manche Namen unmöglich zu matchen (Singlename: “Otto” oder “Döbele” tauchen zu oft auf)
Gender	Neu werden Geschlechter erfasst	
Assigned Roles	Genannte Rollen von Personen im Doc	Läuft, wird aber noch nicht mit Organisationen, sondern nur Personen verbunden **unverädert**
Daten	Kalender-Daten im dd.mm.yyyy Format. Wenn in XML notiert auch “Letzten Samstag” als Datum.	Klappt top
Events	Zusammenfassungen von LLM	Wurde gerade erst begonnen. In Erprobung
Letter-Metadata	Empfänger und Autor, durch regex matching und XML Tags	Klappt ganz ok, Author ist einfacher als Recipient, weil Regex-Formeln auf kaum veränderten Abschlussgruss basiert, (“viele Grüsse”, “dein”, “HH”, etc.) Recipient ist sehr viel schwerer, weil nicht einheitliche Anschrift.
Organisationen	Umfassendes Matching gegen bekannte Organisationen in GT_CSV	Klappt
Orte	Klappt gut, nur Schwierigkeiten bei Orten wie “im Lokal”	Klappt
Document_type	(Brief/Postkarte, Protokoll etc. Darauf basierend andere Anforderungen an JSON	Klappt

Modul	Kurzbeschrieb	Fehler/Fehlerquellen
unmatched_logger	alles, was nicht in der Groundtruth gematched wird wird gespeichert, um es manuell nachzuarbeiten	Aktuell grösste Baustelle, und essentiell für Erweiterung der CSV-GT **In Ordnung**

Aufgetauchte Fragen zur Quellenarbeit / Datenbasis

- **Extreme Heterogenität macht weiterhin Probleme, Ziel alles abzubilden wird unrealistisch.**

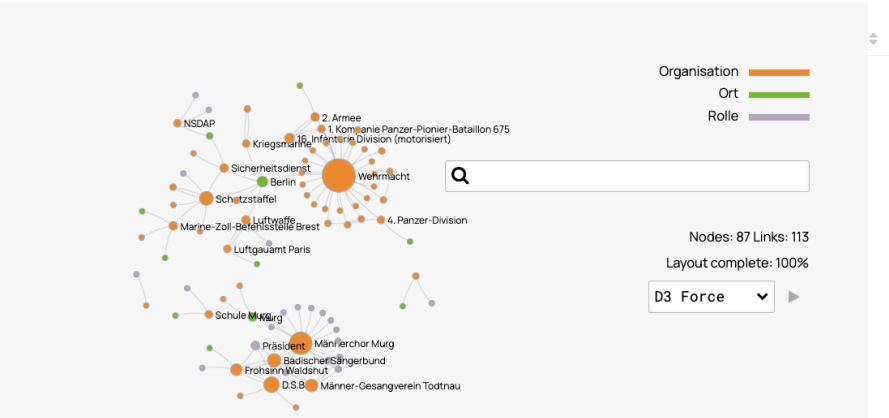
Offene Fragen & Herausforderungen bei der Arbeit an der Masterarbeit

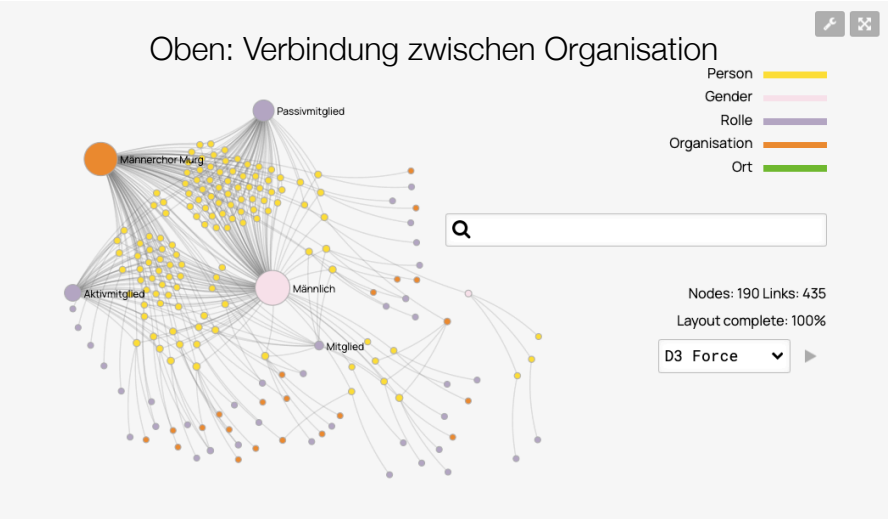
Transkribus:

- Ursprünglich abgeschlossen, aber Probleme beim Export. Tabellen (Sonderform in Transkribus) werden als leere XML exportiert. Die sind allerdings essentiell für Weiterverarbeitung. Problemlösung aktuell vertagt, Pro-Account für Transkribus ist besorgt.

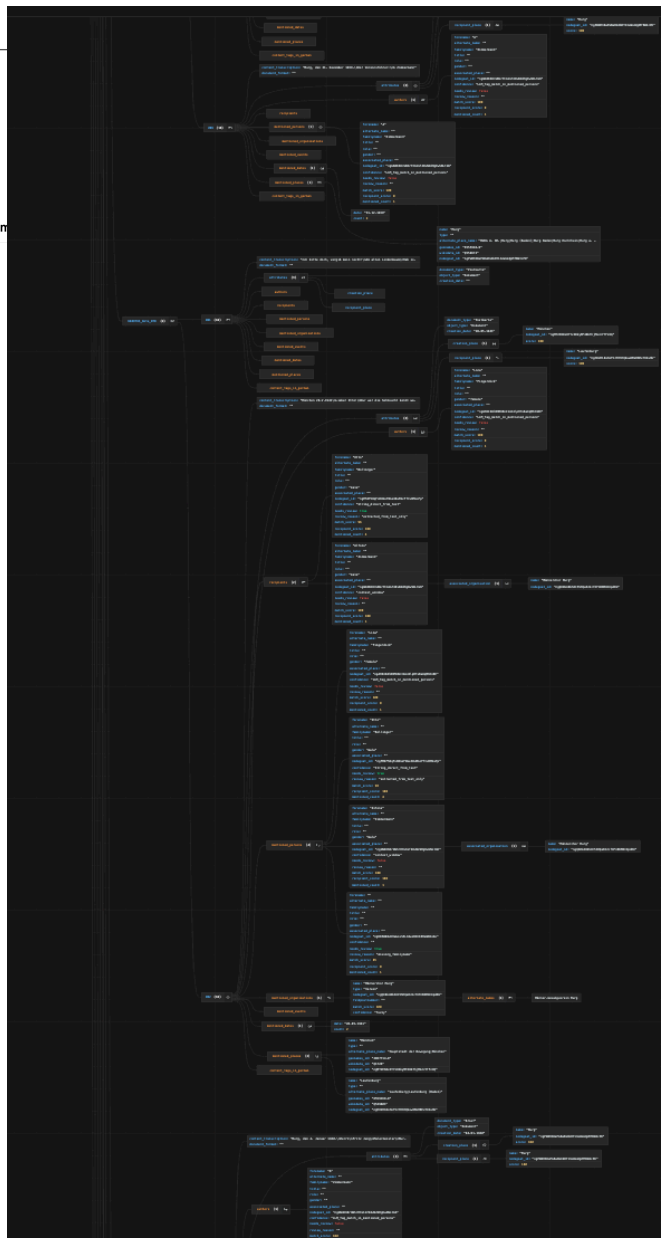
  —> nicht gelöst, Datenset verkleinern

Beispielbilder aktuller Stand:





Oben:Verbindung zwischen Personen



Visualisierung Total_Json mit allen entities