

# D<sup>4</sup> – Dokumente, Diplomatie, Daten und Disambiguierung

## Kontext:

Die automatische Analyse historischer Dokumente ist ein zentraler Bestandteil moderner digitaler geisteswissenschaftlicher Forschung. Projekte der Forschungsstelle Dodis – Diplomatische Dokumente der Schweiz, ein Institut der Schweizerischen Akademie der Geistes- und Sozialwissenschaften, nutzen strukturierte Dokumentformate wie TEI-XML, um historische Quellen maschinenlesbar aufzubereiten und weiterzuverarbeiten. Ein wichtiger Verarbeitungsschritt besteht dabei in der automatischen Erkennung und Verknüpfung von benannten Entitäten (Named Entity Recognition und Entity Linking), um Personen, Orte oder Organisationen eindeutig identifizieren und mit projektspezifischen Wissensbasen verknüpfen zu können.

Während verschiedene Entity-Linking-Lösungen existieren, konzentrieren sich vorwiegend auf allgemeine Wissensbasen wie Wikidata. Diese Ansätze sind für spezifische historische Projekte jedoch oft nur eingeschränkt geeignet, da projektspezifische Entitäten, Metadaten und historische Besonderheiten nicht ausreichend berücksichtigt werden. Zudem sind mehrere bestehende Lösungen veraltet oder nur eingeschränkt gewartet.

Für Dodis soll ein EntityLinker entwickelt werden, der aus Text ein TEI-XML Generiert und die identifizierten Entities (Personen/Orte/Organisationen) mit der DB von Dodis abgleicht. Die würde den Prozess der Indexierung von neuen Dokumenten beschleunigen und den historischen Forschungsprozess vereinfachen.

## Ziel:

Ziel dieses Projekts ist die Konzeption und Implementierung eines projektspezifischen Entity-Linking-Systems für historische Dokumente am Beispiel von Dodis. Dabei soll ein Custom EntityLinker entwickelt werden, der TEI-XML-Dokumente als Trainingsgrundlage nutzt und erkannte Entitäten automatisch mit einer projektspezifischen Wissensbasis verknüpft.

Das Projekt bietet eine Vielzahl von Forschungs- und Entwicklungsoptionen, etwa im Bereich:

- Natural Language Processing für diplomatische Texte
- Entwicklung projektspezifischer Entity-Linking-Strategien
- Verarbeitung und Modellierung strukturierter TEI-Dokumente
- Optimierung von Kandidatengenerierung und Matching-Verfahren
- Evaluation von Entity-Linking-Systemen im historischen Kontext

## Technologie

Je nach Projektfokus und Vorkenntnissen der Studierenden kommen unter anderem folgende Technologien zum Einsatz:

- NLP Frameworks: spaCy
- Datenquellen: TEI-XML-Dokumente und projektspezifische Metadaten
- Persistenz: SQLite oder vergleichbare Datenbanksysteme
- Machine-Learning-Verfahren: tf-idf-basierte Ansätze oder vergleichbare Kandidatengenerierungsmethoden
- Versionsverwaltung: Git

## Charakter des Projekts

Das Projekt verbindet praxisorientierte Softwareentwicklung mit aktueller Forschung im Bereich Natural Language Processing und Digital Humanities. Es bietet sowohl algorithmische als auch datenmodellierungsbezogene Herausforderungen und erlaubt die Entwicklung übertragbarer Methoden für die automatische Analyse historischer Dokumente. Teilaspekte des Projekts können je nach Interesse vertieft oder erweitert werden.

## Beispiel Projekte

Es gibt mehrere Projekte, die einen EntityLinker für Wikidata oder Wikipedia entwickelt haben. Einige funktionieren noch andere sind deprecated oder nur mit veralteten Spacy Versionen nutzbar.

- <https://github.com/egerber/spaCy-entity-linker> Eigenständiges Projekt um Daten mit Wikidata zu verknüpfen
- <https://github.com/explosion/wikid>  
Lädt Wikidata Datensatz herunter und speichert interessante Entities in einer sqlite DB
- <https://github.com/explosion/projects/tree/v3/benchmarks/nel>  
Benchmark/Pipeline, der die Sqlite db von wikid nutzt und das Training orchestriert
- <https://microsoft.github.io/spacy-ann-linker/>  
Related: ein tf-idf Ansatz von Microsoft.
- Dodis hat einen rudimentären Prototype eines EntityLinkers. Dieser funktioniert aber noch nicht zuverlässig.

## Projektbeteiligte

Daten und Infrastruktur: Diplomatische Dokumente der Schweiz (Dodis) [www.dodis.ch](http://www.dodis.ch)

Kundensicht: Merlin Streilein und Tobias Steiner (UniBe), Maurizio Rossi (Dodis)  
{merlin.streilein, tobias.steiner}@unibe.ch