Ökonomien des Raums: Ein historisches Findmittel digital denken

Hodel, Tobias
tobias.hodel@unibe.ch
Universität Bern, Walter Benjamin Kolleg, Schweiz
ORCID: 0000-0002-2071-6407

Burkart, Lucas
lucas.burkart@unibas.ch
Universität Basel, Departement Geschichte, Schweiz
ORCID: 0000-0002-9011-5113

Hitz, Benjamin
benjamin.hitz@unibas.ch
Universität Basel, Departement Geschichte, Schweiz
ORCID: 0000-0002-3208-4881

Aeby, Jonas
jonas.aeby@unibas.ch
Universität Basel, Departement Geschichte, Schweiz

Prada Ziegler, Ismail
ismail.prada@unibe.ch
Universität Bern, Walter Benjamin Kolleg, Schweiz;
Universität Basel, Departement Geschichte, Schweiz
ORCID: 0000-0003-4229-8688

Vonwiller, Aline
a.vonwiller@unibas.ch
Universität Basel, Departement Geschichte, Schweiz

Die Aufbereitung historischer Daten nicht zuletzt aus historischen Findmitteln ist eine enorm gewinnbringende, jedoch auch komplexe Angelegenheit. Dabei werden Fragen zur Aufbereitung und Auswertung mit quellenkritischen Fragen vermischt. Und wenn zusätzlich das Findmittel selbst ein umfangreicher Bestand ist, müssen Unsicherheitsfaktoren, die durch die Anwendung maschineller Lernverfahren entstehen, minimiert, berücksichtigt und adressiert werden. Findmittel sind Hilfsmittel, um sich in den Beständen von Archiven zu orientieren. Häufig konzentrieren sich Findmittel auf bestimmte Bereiche von Archivbeständen.

Alle diese Herausforderungen finden sich kondensiert im Projekt “Ökonomien des Raums”, das am Beispiel der Stadt Basel das Wirtschaften mit dem städtischen Grundbesitz untersucht, das aus der ganzen Breite der schriftlichen Überlieferung im Staatsarchiv Basel-Stadt rekonstruiert wird. Darunter fallen Verzeichnisse zu Pfändungen, Verkaufsurkunden und eine Vielzahl weiterer schriftlich überlieferter Dokumente. Greifbar wird die Überlieferung durch den Einsatz besagter *machine learning* Verfahren, sowie computergestützter Auswertungsverfahren, mit denen das Historische Grundbuch der Stadt Basel (HGB) für die digitale Nutzung aufbereitet wird.

Das Projekt und damit die Repräsentation als Poster reiht sich ein in die Ansätze zur Massenaufbereitung und Informationsextraktion, wobei der Fokus auf die Datenmodellierung (insbesondere mit Blick auf Geodaten) und auf die Extraktion und Kategorisierung von Ereignissen gelegt wird. Besonders der zweite Schritt nutzt dabei Ansätze des *deep learning* und *token*-Vektorisierung.

Das Historische Grundbuch der Stadt Basel (HGB)

Um 1900 wurden alle damals greifbaren Archivalien für das HGB in knappen Quellenauszügen exzerpiert und handschriftlich auf Zettelkarten notiert. Eine Karte steht dabei typischerweise für eine Transaktion oder eine Eintragung in einem Quellenstück. Der Inhalt des Quellenstücks wurde exzerpiert und meist in der Quellensprache auf der Zettelkarte abgeschrieben. Die Form des Exzerpts soll und kann in der digitalen Form nicht rückgängig gemacht werden. Dieser Umstand und das HGB als Findmittel ist für die Auswertungsstrategien und -methoden zentral. Damit können wir indirekt auf die zumeist integral erhaltenen Quellenbestände zurückgreifen, die wiederum über stichprobenartige Überprüfungen ausgewählter Bestände analysiert werden können.

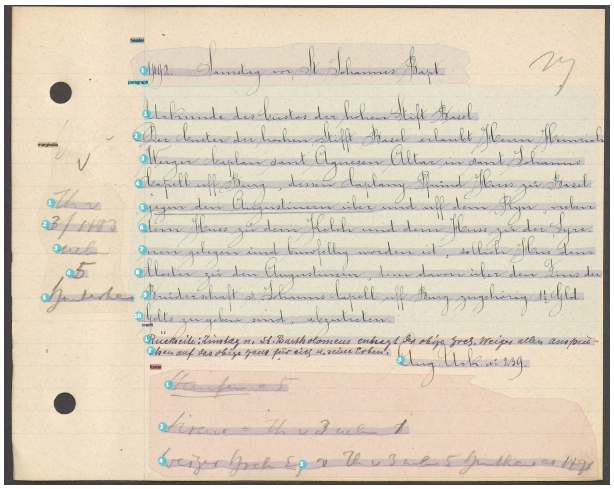

Abb. 1: Beispiel einer semantisch analysierten Seite. Screenshot aus Transkribus.

Die Strukturierung der Information auf diesen Karteikarten ist nicht «neutral», sie folgt den Interessen der Urheber des HGB ebenso wie Konventionen der Informationserfassung; somit besteht für die computergestützte Auswertung

strukturell eine Differenz zwischen den digital verfügbaren Informationen und einer qualitativen Lektüre der Quellen.

Auf sprachlicher Ebene hingegen stellt sich das Problem eines spezifischen Einflusses durch die Urheber weniger, weil die Angaben meist als Kurzzitat in den Begriffen und Formulierungen der Quellen selbst erfolgt sind. Wenn auch aus dem Kontext der Dokumente gerissen, ermöglicht dies dennoch eine Analyse der Quellensprache, ihrer Begrifflichkeit und Semantiken.

Datenmodell und Workflow

Das Poster diskutiert das Projekt, wobei Datenmodell und Workflows zur Aufbereitung von rund 120’000 für uns relevanten digitalisierten Bilddateien zentral präsentiert werden. Dabei wird großer Wert auf Reproduzierbarkeit und die Möglichkeit zum wiederholten Prozessieren der genutzten Algorithmen gelegt.

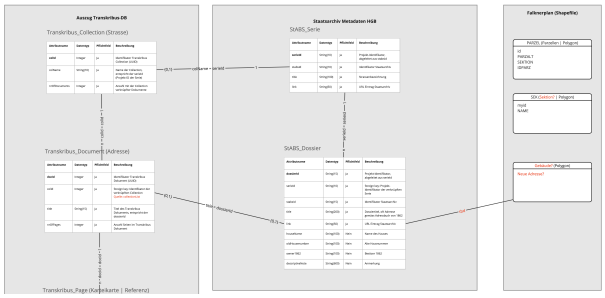

Abb. 2: Auszug aus der Modellbeschreibung der relationalen Projektdatenbank, umgesetzt in PostgreSQL. Screenshot: Projektdokumentation.

Das Datenmodell bildet aus unterschiedlichen Perspektiven und mit unterschiedlicher Granularität die vorhandenen Informationen ab. Einerseits wird auf archivische Metadaten (Signatur und Adressen als Strings, bezogen von einem SPARQL-Endpoint) und bereits vorliegende Geodaten (in Form von “Shapefiles”) zurückgegriffen, andererseits wird aus den gescannten Dokumenten in einem mehrstufigen Prozess Informationen extrahiert: Beginnend auf der Stufe “Sammlung” (Strasse), über “Einzeldokument” (Adresse), zur “Seite” (Karteikarte mit typischerweise einem Ereignis) und schliesslich der Textregion mit den zentralen Informationen (in einem “header” und einem “paragraph”).

Auf technischer Ebene bedeutet dies, dass die Dokumentenmassen in semantisch angereicherte Textregionen aufgeteilt werden (Quirós 2017), bevor eine Texterkennung der Handschriften durchgeführt wurde. Für beides, Segmentierung und Texterkennung, wurden spezifische Modelle erstellt, die eine zuverlässige Identifikation ermöglichen. Insbesondere die Texterkennung mit PyLaia (Puigcerver [2017] 2022), die mit mehr als zehn verschiedenen Händen umgehen musste (Hodel u. a. 2021; Hodel 2023; Pinche 2023), erzielt für ein grosses Modell überzeugende Fehlerraten im Bereich von 4% auf einem Testset. Die Resultate werden mit 3,5% noch besser, wenn nur auf die zentralen Bereiche (die angesprochenen “header” und “paragraph”) fokussiert wird. Die mit vier Prozent Character Error Rate erzeugten Texte sind für die darauf folgenden Auswertungsschritte, insbesondere die Extraktion von benannten Entitäten (Personen, Orte, Organisationen), deren Relationen und genannte Events mit Hilfe spezifischer Sprachmodelle ausreichend (Torres Aguilar und Stutzmann 2021; Cafiero u. a. 2021; Hodel, Prada Ziegler, und Schneider 2023).

Datenvisualisierungen

Basierend auf den extrahierten Daten und der Kombination mit den identifizierbaren Grundstücken, wird es möglich, nach Begriffen (im Projekt “Eventtypen”) zu suchen und diese auf Karten zu visualisieren, etwa unter Berücksichtigung von Zeitschnitten, wie in Abb. 3 demonstriert.

Die Analyse von Events und Transaktionen bedingt Erkenntnisse zum normativen und semantischen Wandel des Liegenschaftsmarktes. Nur so können Überlieferungslücken, sprachliche Veränderungen und historischer Wandel auseinandergehalten und entsprechend visualisiert werden. Referenz ist dabei der HGB-Bestand: wie dicht ist die Überlieferung in einem Zeitraum und wie gross der Anteil von kategorisierbaren Karteikarten in der Erkennung von Events.

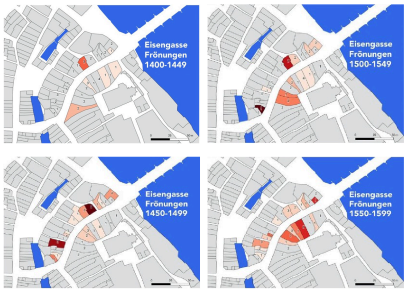

Abb. 3: Beispiel einer visuellen Auswertung mit der Filterung «Frönungen» (Event) gekoppelt mit der Häufigkeit des Auftretens über Zeitschnitte.

Zwischenresultate

Obwohl sich das Projekt noch in einer frühen Phase befindet und hinsichtlich der Auswertung weitere Verfeinerungen in den kommenden Jahren erwartet werden dürfen, wagen wir erste Aussagen.

Mit dem erarbeiteten Datenmaterial wird der Faktormarkt Boden im Kontext der dynamischen urbanen Ökonomien des Spätmittelalters und dessen formelle und informelle Strukturierungen als Voraussetzung seiner Funktionsweise analysiert.

Die Datenbasis eines digitalen HGB bietet die Chance, Untersuchungen des vormodernen städtischen Immobilien-

marktes in synchroner und diachroner Perspektive einerseits, in räumlicher Skalierbarkeit von einzelnen Häusern bis zum gesamten Stadtraum andererseits, durchzuführen. Sie bietet die Grundlage für eine Heuristik ausgesprochen dichter Überlieferung, mit der in einer Zeit-Raum-Matrix Akteure, Praktiken und Vokabularien der «Ökonomien des Raums» variabel analysiert werden können.

Gleichzeitig zeigen die eingesetzten Methoden, dass auch für grosse Datensätze die maschinellen Lernverfahren einen reifen Status erreicht haben. Die Auswertung von Massenquellen und komplexer Findmittel kann also in Angriff genommen werden.

Bibliographie

Cafiero, Florian, Thibault Clérice, Paul Fièvre, Simon Gabay, und Jean-Baptiste Camps. 2021. „Corpus and Models for Lemmatisation and POS-tagging of Classical French Theatre“. *Journal of Data Mining & Digital Humanities* 2021 (Februar). https://doi.org/10.46298/jdmdh.6485.

Hodel, Tobias. 2023. „Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik“. *Historische Zeitschrift* 316 (1): 151–80. https://doi.org/10.1515/hzhz-2023-0006.

Hodel, Tobias, Ismail Prada Ziegler, und Christa Schneider. 2023. „Pre-Modern Data: Applying Language Modeling and Named Entity Recognition on Criminal Records in the City of Bern“. Graz, Austria, Juni 30. https://doi.org/10.5281/zenodo.8107616.

Hodel, Tobias, David Schoch, Christa Schneider, und Jake Purcell. 2021. „General Models for Handwritten Text Recognition: Feasibility and State-of-the Art. German Kurrent as an Example“. *Journal of Open Humanities Data*, Journal of Open Humanities Data, 7. https://doi.org/10.5334/johd.46.

Pinche, Ariane. 2023. „Generic HTR Models for Medieval Manuscripts The CREMMALab Project“. https://hal.science/hal-03837519.

Puigcerver, Joan. (2017) 2022. „PyLaia“. Python. https://github.com/jpuigcerver/PyLaia.

Quirós, Lorenzo. 2017. „P2PaLA: page to PAGE layout analysis toolkit“. https://github.com/lquirosd/P2PaLA.

Torres Aguilar, Sergio, und Dominique Stutzmann. 2021. „Named Entity Recognition for French medieval charters“. In *Workshop on Natural Language Processing for Digital Humanities*. Workshop on Natural Language Processing for Digital Humanities Proceedings of the Workshop. Helsinki, Finland. https://hal.archives-ouvertes.fr/hal-03503055.