universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

Reconstructing Administrative Structures of the “Hof- und Staatsschematismus” of the Habsburg Monarchy: An Evaluation of Automatically Extracted Layout Information

verfasst von | submitted by

Bernhard Ortbauer Bakk.rer.nat. BSc BA M.Sc.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 647
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Digital Humanities
Betreut von | Supervisor:	Mag. Dr. Thomas J.J. Wallnig Privatdoz. MAS
Mitbetreut von | Co-Supervisor:	Mag. Dr. Wolfgang Göderle

German Abstract

Der „Hof- und Staatsschematismus" („Hof- und Staatshandbuch") der Habsburgermonarchie ist ein von 1702 bis 1918 erschienener Almanach, in dem eine beträchtliche Anzahl von Hof- und Staatsbeamten systematisch verzeichnet wurde. Die ursprünglich als jährliches Periodikum konzipierte Serie weist zwar mehrere Lücken auf, über die Österreichische Nationalbibliothek können aber immerhin 154 Bände mit jeweils zwischen 150 und 1900 Seiten frei zugänglich online abgerufen werden.[1] Grob geschätzt enthalten diese Jahrgänge über 150.000 verschiedene Personen, was den Datensatz zu einer wertvollen geschichtswissenschaftlichen Quelle macht. Ein kooperatives Forschungsteam der Universität Graz und der TU Graz widmet sich aktuell der automatisierten, KI-gestützten Informationsextraktion aus dem Schematismus. Das im Zuge dieser Bemühungen trainierte Machine-Learning-Modell erkennt Layouteigenschaften der gescannten Druckseiten mit hoher Genauigkeit.

Ziel der vorliegenden Masterarbeit ist es, anhand der Ausgabe von 1878 zu evaluieren, ob die gewonnenen Informationen für eine Rekonstruktion der im Druckwerk abgebildeten hierarchischen Ämtergliederung geeignet und ausreichend sind. Die zu diesem Zweck durchgeführte komparative Analyse zweier Rekonstruktionsheuristiken zeigt, dass weder das lückenlose Wissen um die Abfolge von Überschriften und darunter gelisteten Personen noch die überwiegend zuverlässige automatische Klassifizierung der Überschriften in sechs Kategorien ausreicht, um die hierarchische Ämtergliederung vollständig abzuleiten. Aufgrund des negativen Befundes wird das Druckwerk in der Folge einer eingehenden Analyse unterzogen, um weitere extrahierbare Layout-Elemente zu identifizieren und deren Konsistenz zu evaluieren. Die Untersuchungen fördern unter anderem gedruckte Trennlinien als zuverlässige Markierungen für Abschnittsgrenzen zutage. Dennoch muss konstatiert werden, dass nur eine detaillierte Disambiguierung von Überschriftenstilen eine vollständige Entschlüsselung der Ämterhierarchie auf Basis von Layouteigenschaften ermöglicht. Im letzten Abschnitt der Arbeit wird daher vorgeschlagen, die Bildausschnitte einzelner Überschriften durch Clusteranalyse oder Klassifizierungsalgorithmen zu separieren und so Möglichkeiten für eine noch detailliertere Aufschlüsselung verschiedener Überschriftenstile zu erhalten.

[1] https://alex.onb.ac.at/static_tables/shb.htm (last accessed: 21.07.2025)

English Abstract

The "Hof- und Staatsschematismus" (“Hof- und Staatshandbuch“) of the Habsburg monarchy is an almanac published from 1702 to 1918 that systematically recorded a considerable number of court and state officials. While the series, which was originally designed as an annual publication, has several gaps, 154 volumes with 150 to 1900 pages each are freely accessible through the web-appearance of the Austrian National Library.[2] These volumes contain an estimated 150,000 different individuals, making the dataset a valuable source for social history research. A collaborative research team from the University of Graz and Graz University of Technology is currently working on automated, AI-supported information extraction from the Schematismus. The machine learning model trained for this purpose recognizes layout properties of the scanned printed pages with high accuracy.

The goal of this master's thesis is to evaluate whether the extracted information is suitable and sufficient for reconstructing the hierarchical office structure depicted in the printed work based on a data sample taken from the 1878 edition. A comparative analysis of two reconstruction heuristics shows that neither complete knowledge of the sequence of headings and listed individuals nor the largely reliable automatic classification of headings into six distinct categories is sufficient to fully derive the hierarchical office structure. Due to this negative finding, the printed work is subsequently subjected to detailed analysis to identify additionally extractable layout elements and evaluate their consistency. The investigations reveal printed separation lines as reliable markers for section boundaries. Nevertheless, it must be concluded that only detailed disambiguation of heading styles enables complete decoding of the office hierarchy based on layout properties. The final section of the thesis, therefore, proposes strategies to enrich the information on image segments of individual headings through cluster analysis or classification algorithms in order to obtain a more detailed breakdown of different heading styles.

[2] https://alex.onb.ac.at/static_tables/shb.htm (last accessed: 21.07.2025)

Danksagung

Ich danke Dr. Thomas Wallnig herzlich für seine engagierte, hilfreiche und ehrliche Betreuung. Dr. Wolfgang Göderle danke ich nicht nur für die Betreuung dieser Arbeit, sondern vor allem für sein mir entgegengebrachtes Vertrauen und den langen freundschaftlichen Austausch, der mich schon durch meine Bachelorarbeit begleitet hat und im Rahmen des Schematismusprojektes hoffentlich noch eine längere Fortsetzung findet.

Den größten Dank schulde ich meiner Familie und meiner Freundin Silvia für ihre bedingungslose Unterstützung und ihre Geduld.

Contents

1 Introduction

Digital data is of superior significance in today's world. We sometimes perceive it as so defining for our time that we might forget that structured data was also produced before the digital age. The "Hof- und Staatsschematismus" of the Habsburg monarchy is but one impressive example of the extent registers from the 18th and 19th centuries could reach. Over more than 200 years 154 updated and ever-growing editions listed the officers of the Habsburg court and state administration, from the highest ranks down to laundry women or middle school headmasters. The huge amount of information makes the Schematismus a very rich source for historical scholars, but also one that is very difficult to access in its entirety.

An ongoing joint research project at the University of Graz and Graz University of Technology aims at automatically transferring the scanned editions of the Schematismus into a digital database to make the hierarchies of Habsburg bureaucracy digitally searchable and open to quantitative and qualitative analysis. Conceptually, the printed Schematismus may be considered a database already. The information is stored in a structured manner using the technologies of the 18th and 19th century. This information must be transformed into a digital format without introducing errors. Essentially, the printed Schematismus functions as a tangible representation of the organizational structure of the Habsburg court and state administration. The transfer of this meticulously structured printed format into a digital database necessitates the precise identification of individual layout elements and a faithful reconstruction of their hierarchical relationships.

Automatic layout detection based on machine learning has proven to be a promising way to master this complex challenge. The research group led by Wolfgang Göderle and Roman Kern successfully employed computer vision models to split the textual layout of individual pages into several distinct layout classes, including a very reliable separation between headings and paragraphs. This approach mimics the structural logic of the printed Schematismus but with a reduced level of detail compared to the multitude of differently styled headings in the original source. Whether the hierarchical office structure can be recreated from this information remains an open question and serves as impetus for the present master's thesis.

This reconstruction process represents an important methodological bridge between the extracted raw data and the creation of a meaningful prosopographical database that

can serve historical scholarship. Picking up the structured data output of the computer vision model for layout extraction, this thesis aims at evaluating approaches towards reconstructing the hierarchical organizational structure represented in the Schematismus. There is general expectation that large parts of the hierarchical relationships can, in fact, be reconstructed solely based on visual criteria, since the highly fragmented layout of the Schematismus operates with a multitude of differently styled, nested headings to demarcate organizational boundaries. The evaluations in this thesis should explore possibilities and boundaries of such reconstruction approaches.

The final structure of the present thesis was shaped by the different information requirements that are tied to this evaluation task. It is necessary to determine and understand which explicit and implicit forms of representation were used to indicate hierarchical and structural boundaries in the printed work. Furthermore, it is advisable to have a general grasp of the possibilities and limitations of the technologies used for data extraction and to contemplate the common practices within prosopographical research regarding data modeling and database design.

All of these context related contents, along with the concrete evaluations, are distributed over seven chapters. Following this introduction, chapter 2 is supposed to offer a short presentation of the historical source under consideration. The "Hof- und Staatsschematismus" will be contextualized historically and analyzed with regards to significant changes of its visual appearance that occurred during its long lifespan. The visual appearance of the printed source is of major significance for the layout detection approaches of the Schematismus Project.

The theoretical chapters of this master's thesis are supposed to draw broad lines. They should give an overview and prepare readers for the later chapters on very specific practical problems. In this vein, chapter 3 serves as an introduction into common data modeling practices of prosopographical research projects. The chapter summarizes the evolution of the research field and its methodological conventions. The explanation of data modelling practices of digital prosopographies is guided by examples. A brief presentation of prosopographical projects of the Habsburg court and state is used to readjust the focus of the thesis to the "Hof- und Staatsschematismus" and gives a first indication regarding the potential value of a digital Schematismus database.

The overarching topic of chapter 4 is the automated layout extraction. First, a wide-angle overview is given of the computer vision applications, which form the technological

basis of extraction. This is followed by a detailed analysis of the data representation of the printed source, in order to understand the different entities that need to be recognized and their respective representation in the print. Based on this subsection and subsequent to it, the annotation schema that was used to produce training data for the extraction model, and which predefines the anticipated output is described. Understanding this upstream part of the research process is vital not only for a conclusive picture of the information in the dataset, but also for comprehending the translational process that has to take place between computational humanists and computer and data scientists in such a complex research scenario. The chapter is concluded by the introduction of a simple schematic data model for the data representation of the Schematismus.

In Chapter 5 a rule-based approach is developed to manipulate the extracted data in a way that should allow the reproduction of hierarchical relationships reflected in the printed source and facilitate translation into the preliminary data model introduced earlier. The reproduction approach builds on a deconstructed logic of sequential positioning of extracted elements, which is consequently evaluated in comparison to a similar reconstruction effort carried out by the research group from Graz.

Both approaches deliver promising results but are not able to deliver a conclusive hierarchical structure. Therefore, chapter 6 initiates a systematic investigation to find further extractable relevant information within the source. Based on the findings, different strategies to achieve a more detailed classification of heading styles are discussed. The strategies encompass considerations regarding the post-processing of already extracted information as well as additional extractions. Chapter 7 ultimately holds the concluding remarks of the thesis.

Before entering the deliberations of the main part a quick overview of frequently used terms and definitions is added. Since the thesis focuses on very specific problems regarding a very specific historical source, this measure is thought to improve the fluidity of the text and to serve as reference for readers to return to in case of uncertainty.

1.1 Terms and Definitions

Schematismus:

In this thesis the term Schematismus refers to a specific series of a state almanac of the Habsburg court and state. Editions from this series were published between 1702 and 1918.

The scanned images of this series are publicly accessible via the homepage of the Austrian National Library (ÖNB).[3] The ÖNB uses "Staatshandbuch" as umbrella term for the entire series and divides it into four major subsections (1702 – 1806, 1807 – 1843. 1844 – 1873, 1874 – 1918). The publication had in fact several official names during its lifespan, for example "Kayserlicher Und Königlicher Wie auch Ertz-Hertzoglicher Und Dero Residentz-Stadt Wien Staats- und Stands- Calender", "Hof- und Staats-Schematismus des österreichischen Kaiserthums" or "Hof- und Staatshandbuch der Österreichisch-Ungarischen Monarchie"[4] Most commonly the source is referred to as "Staatshandbuch" or as is the case throughout this thesis as "Schematismus".

Schematismus Project:

The term "Schematismus Project" is used throughout this thesis to refer to a joint-research project at the University of Graz and Graz University of Technology. The research group led by Wolfgang Göderle and Roman Kern aims at automatically extracting and compiling information from the Schematismus. The project has so far produced promising results in the field of automated layout detection and has been granted funding for a conclusive digitization of the entire Schematismus series. The first phase of the project is aimed at the editions between 1876 and 1918.

Layout Model:

The term "Layout Model" refers to the machine learning model that was trained by David Fleischhacker, a member of the Schematismus Project's research team. A recent publication outlines the innovative training process of the model and showcases the capabilities with regards to layout detection.[5] Because of the fluidity of the field of machine learning, the technological architecture of the model is constantly evolving. In this thesis the term "Layout Model" most of the times refers either to its latest iteration or to the one used to produce the dataset that serves as input for the practical evaluations of the present thesis in July 2024.

---

[3] https://alex.onb.ac.at/static_tables/shb.htm (last accessed: 08.05.2025).
[4] Ibid.
[5] FLEISCHHACKER David / GOEDERLE Wolfgang et al., Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach 2024. Online reference: arxiv.org/pdf/2401.07787, (last accessed: 08.02.2025).

2 The Hof- und Staatsschematismus of the Habsburg Monarchy

Initially, the attention will turn to the historical source, which is at the core of the Schematismus Project. The Schematismus of the Habsburg monarchy was a close to yearly published record of individuals employed at the Habsburg court and state, structured along the lines of administrative subdivisions. The next few pages will explore the genesis and historical evolution of the Schematismus. This wide-angle overview is followed by a review of changes regarding layout and visual appearance of the Schematismus over the 200 years it was published.

2.1 Genesis and Evolution of the Schematismus Series

The keeping of so-called "Hofordnungen"[6] goes back at least as far as the Middle Ages.[7] Long before enlightened monarchies invented and expanded civil services in the modern sense of the word, rulers saw the need to meticulously document the titles and offices they awarded and the staff they paid – not least to track and limit expenditure.[8] From the late 17th century onwards, court and state almanacs ("Amtskalender", "Amtshandbücher", "Amtsschematismen"),[9] which followed on from the "Hofordnungen" in terms of content, but which were published periodically under official sanction, emerged as a genre in their own right.[10] The genre became established in most large and small dominions of the Holy Roman Empire over the course of the 18th century. Bauer compiled a register of all such almanacs within the empire. He noted 3,036 publications of court and state almanacs for the various dominions in the 18th century.[11] Surely these editions vary widely in size and

[6] Hofordnungen or „Ordonnances de l'hotel" were registers that described the offices of the court and listed the personnel for each office along with respective service period and remuneration. cf. NOFLATSCHER Heinz, Ordonnances de l'hôtel, Hofstaatsverzeichnisse, Hof- und Staatskalender. In: Josef PAUSER / Martin SCHEUTZ / Thomas WINKELBAUER (Eds.) Quellenkunde der Habsburgermonarchie. (16. - 18. Jahrhundert) ; ein exemplarisches Handbuch (= Mitteilungen des Instituts für Österreichische Geschichtsforschung Ergänzungsband 44). Wien – München: Oldenbourg, 2004, pp. 59–75, here: p. 63.

[7] WIDDER Ellen, Hofordnungen. In: Handbuch Höfe und Residenzen im spätmittelalterlichen Reich - Band15.III (= Residenzenforschung 15,3). Ostfildern: Thorbecke, 2007, pp. 391–408, here: pp. 391–392.

[8] ANDERMANN Kurt, Pragmatische Schriftlichkeit. In: Handbuch Höfe und Residenzen im spätmittelalterlichen Reich - Band15.III, pp. 37–60, here: pp. 48–50.

[9] In this thesis the genre is collectively referred to as court and state almanac(s), a characterization that is also used by Volker Bauer. cf. BAUER Volker, Territoriale Amtskalender und Amtshandbücher im Alten Reich. In: Rechtsgeschichte - Legal History 2002, 01, 2002, pp. 71–89, Abstract.

[10] Ibid., pp. 71–73.

[11] Ibid., p. 73.

detail, but one can imagine how big a potential for socioeconomical and prosopographical historical research these sources could hold.

The genesis of court and state almanacs seems to vary regionally, but the function that the rulers saw in the publication of such registers of offices was primarily a ceremonial and representative one.[12] The almanacs were used to demonstrate the size and structure of the court and in the case of bigger houses such as the House of Habsburg the almanac could serve as an example for other sovereigns to shape their court accordingly. Although these publications were initially very commonly printed and edited by private enterprises, their contents and distribution were always under tight governmental control and publication was only possible under official license.[13]

For the Habsburg court and state, several such almanacs existed, among them a very comprehensive publication, which was edited from 1702 onwards until 1918. Initially published under the name "Kayserlicher Und Königlicher Wie auch Ertz-Hertzoglicher Und Dero Residentz-Stadt Wien Staats- und Stands- Calender", the series quickly became known as the "Hof- und Staatsschematismus" or the "Schematismus" in short.

As one of the most extensive court and state almanacs in the territory of the Holy Roman Empire, the Schematismus contains by name all administrative personnel of the Habsburg court and state down to comparatively low ranks. The more than 200 years during which the Schematismus was published with some intermissions, have witnessed social and administrative transformations like the rise of the civil service and the emergence of an educated middle class as well as political breaks like the founding of the Austrian Empire in 1804, the dissolution of the Holy Roman Empire in 1806, the revolution of 1848 or the Austro-Hungarian Compromise of 1867.[14] Enabling detailed quantitative and qualitative investigations into the administrative body of the Habsburg bureaucracy on a longitudinal time series is a major goal of the present digitization project.

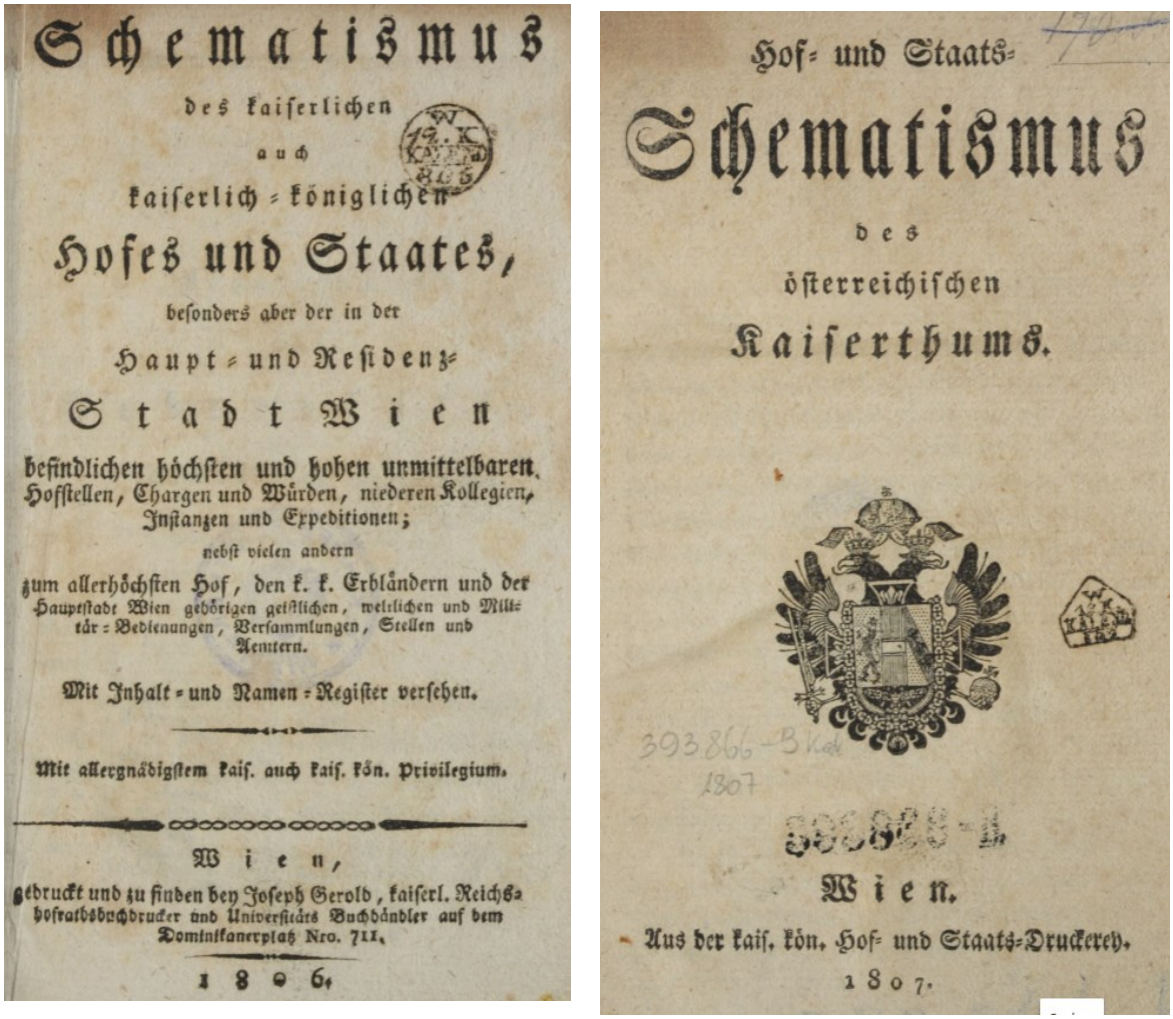
The Schematismus series comprises a total of 153 editions, 78 of which date from the 19th century. Of course, the series was subject to major changes over the entire period. The first edition from 1702, with its 180 pages, looks almost tiny compared to the last 1,907-

[12] NOFLATSCHER, Ordonnances de l'hôtel, Hofstaatsverzeichnisse, Hof- und Staatskalender, 2004, p 59.

[13] BAUER, Territoriale Amtskalender und Amtshandbücher im Alten Reich, 2002, pp. 74–76.

[14] FLEISCHHACKER / GOEDERLE et al., Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach, p. 2.

page edition from 1918. The volumes grew steadily over the years, and the information density varied. In total, the series comprises over 150,000 pages.[15]



Schematismus
des kaiserlichen
auch
kaiserlich-königlichen
Hofes und Staates,
besonders aber der in der
Haupt- und Residenz-
Stadt Wien
befindlichen höchsten und hohen unmittelbaren
Hofstellen, Chargen und Würden, niederen Kollegien,
Instanzen und Expeditionen;
nebst vielen andern
zum allerhöchsten Hof, den k. k. Erbländern und der
Hauptstadt Wien gehörigen geistlichen, weltlichen und Militär-Bedienungen, Versammlungen, Stellen und
Aemtern.
Mit Inhalt- und Namen-Register versehen.
Mit allergnädigstem kais. auch kais. kön. Privilegium.
Wien,
gedruckt und zu finden bey Joseph Gerold, kaiserl. Reichshofrathsbuchdrucker und Universitäts Buchhändler auf dem
Dominikanerplatz Nro. 711.
1806.

Hof- und Staats-
Schematismus
des
österreichischen
Kaiserthums.
Wien.
Aus der kais. kön. Hof- und Staats-Druckerey.
1807.

Figure 1: Comparison between the title pages of the Schematismus from 1806 (left) and 1807 (right).[16]

Until 1806 the Schematismus of the Habsburg court and state was edited by the printer Joseph Gerold, "kaiserlicher Reichshofratsbuchdrucker und Universitäts Buchhändler" (see Figure 1). He wrote in the preface to the 1800 Schematismus:

„... Was mich anbelangt, wird man, hoffe ich, aus allen ersehen, dass ich alle Sorgfallt anwende, diesem unentbehrlichem Buche den möglichsten Grad der Vollkommenheit, und allen Ständen durch diesen [sic!] Schema den vollkommensten Unterricht von dem, was

[15] The figures in this section are based on the estimates Wolfgang Göderle gave in a talk in Graz in February 2024: GÖDERLE Wolfgang, Workshop - Extracting the Hof- und Staatshandbuch. Graz 2024.

[16] Digitized by the ÖNB and publicly accessible under: https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1806&page=1 and https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1807&page=1, (last accessed: 06.01.2025).

sie in diesem Fache zu wissen nöthig haben, zu geben. Der Nutzen dieses Buches ist zu bekannt, und daher einer weitern Empfehlung überflüssig. Ich halte mich für die Mühe hinlänglich belohnt, wenn das Publikum dieses so sehr verbesserte Werk mit gütigen Beyfall beehret. Sollten aber dennoch einige Fehler wider Wissen und Willen eingeschlichen seyn, so wird geziemend gebeten, solche nicht als vorsetzlich zu betrachten; vielmehr hochgeneigtest zu verfügen, daß die Anzeige an gehörigem Orte hievon geschehe, alsdenn wird in nächst folgendem Drucke die Abänderung ganz sicher geschehen. "[17]

"... As far as I am concerned, one will, I hope, see from all of this that I am taking every care to give this indispensable book the greatest possible degree of perfection, and to provide all classes with the most complete instruction in what they need to know in this subject through this scheme. The usefulness of this book is too well known, and therefore no further recommendation is necessary. I consider myself sufficiently rewarded for my efforts if the public honors this much-improved work with gracious applause. Should, however, some errors have crept in against my knowledge and will, it is duly requested that they not be regarded as deliberate; rather, it is most respectfully requested that they be reported in the proper place, as the correction will certainly be made in the next printing."

We must bear in mind that this is the editor's choice of words, but Joseph Gerold confidently assures his readers that his book is an indispensable work, whose usefulness is well known and proven. Whether the "most perfect instruction of what they need to know in this subject" means all the information they "need" or rather they "are entitled to" is a matter of interpretation. However, the fact that the series had already been in existence for almost a hundred years at this time and that the scope and wealth of detail grew with each edition could be an indication that the genre was well established by the time and that there was a lively demand for the annually published Schematismus.

Bauer argues that in the late 18th century the representative character of court and state almanacs gradually gave way to a statistical and controlling function for the state and the public. Originally a medium intended to propagate the might and glamour of the monarch and its court, the almanac eventually turned to evidence, which could be used to

[17] Preface to the Schematismus from 1800. Digitized by the ÖNB and publicly accessible under: https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1800&page=1 (last accessed: 6 January 2025).

criticize rulers for inflated court expenditure.[18] It seems appropriate to assume a similar change in meaning and interpretation for the Schematismus around that time. In any case the edition of the Schematismus experienced several transformations in the first half of the 19th century. Temporally related to the foundation of the Austrian Empire and the dissolution of the Holy Roman Empire, the publication of the Schematismus was transferred to the "k. k. Hof- Und Staatsdruckerei" (Imperial-Royal Court and State Printers) from 1807 onwards. Whether it was due to the Zeitgeist, which desired increased state control over all possible printed works, or because there were too many errors in the private editions, as editor Joseph Gerold submissively admits in his foreword quoted above is not known.

The print series was continued under state responsibility over the next century and still expanded considerably. According to Waltraud Heindl, the civil service by the middle of the 19th century had evolved from its initial dynamics under Joseph II, who had furthered the development of institutional practices that saw the individual officer work its way up from a "Kreisamt" (circle offices), into a rigid and inert state administration, which raised concern and criticism among contemporaries.[19] Paradoxically, the perceived stasis of the apparatus went hand in hand with an inflation of the civil service personnel. While the higher ranks saw only a moderate rise of officers – albeit with a remarkable proportional increase of the higher nobility – the lower ranking servants and interns more than doubled between 1828 and 1846.[20] Especially the increased use of interns seems to reflect that evermore administrative work had to be tackled with very limited budget constraints.[21] The growth of the civil service is reflected in the steep increase of individuals included in the Schematismus volumes during those years.

The years following the revolution of 1848 then brought grave changes to the administrative structure of the monarchy. While on a constitutional level initial concessions regarding representation were quickly revoked and absolute powers of the young Emperor Franz Joseph restored, his leading ministers initiated far-reaching administrative, social and economic reforms to address the need for modernization that had shown itself too

---

[18] BAUER, Territoriale Amtskalender und Amtshandbücher im Alten Reich, 2002, pp. 78–82.

[19] HEINDL Waltraud, Gehorsame Rebellen. Bürokratie und Beamte in Österreich 1780 bis 1848 (=Studien zu Politik und Verwaltung 36). Wien: Böhlau 1990, 16-18; 31-34; 47-52.

[20] Ibid., pp. 134–59.

[21] DEAK John, Forging a multinational state. State making in imperial Austria from the Enlightenment to the First World War (=Stanford studies on Central and Eastern Europe). Stanford, California: Stanford University Press 2015, p. 58.

clearly before and during the revolution.[22] The administrative reforms entailed a restructuring of the administrative units. The provinces became crownlands with a "Statthalter" or "Landespräsident" (governor) as highest regional authority – even if in total dependence of the central bureaucracy.[23] The central government was restructured too.

Figure 2 shows two extracts from the tables of contents for the years 1840 and 1853 – before and after the implementation of the administrative reforms. Before the revolution the "Staats-Conferenz", the "Geheime Cabinett Sr. Majestät des Kaisers" and the "K. K. geh. Haus- Hof- und Staatskanzley" included some of the highest offices. The subdivision of the highest offices in 1853 shows modernization in terms of print, structure and terminology. Following the reforms of the early 1850ies the highest state offices after the emperor were called ministries.

Der Staat.

Erste Abtheilung.

Zweyte Abtheilung.

Hofstellen.

1840

Minister - Konferenz.

Ministerium des kais. Hauses und des Aeusseren.

Ministerium des Inneren.

1853

Figure 2: Comparison of two extracts from the tables of contents of the Schematismus from 1840 (left) and the Schematismus from 1853 (right).[24]

[22] RUMPLER Helmut, Integration und Modernisierung. Der historische Ort des „Neoabsolutismus" in der Geschichte der Habsburgermonarchie. In: Harm-Hinrich BRANDT (Ed.) Der österreichische Neoabsolutismus als Verfassungs- und Verwaltungsproblem. Diskussionen über einen strittigen Epochenbegriff (= Veröffentlichungen der Kommission für Neuere Geschichte Österreichs Band 108). Wien – Köln et al.: Böhlau, 2014, pp. 73–81, here: pp. 75–80.

[23] DEAK, Forging a multinational state, 2015, pp. 112–17.

[24] Digitized by the ÖNB and publicly accessible under: https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1840&page=19 and https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1853&page=9, (last accessed: 06.01.2025).

We would expect all those changes and their entailing personnel reshuffling to be well represented in the Schematismus. Unfortunately, one of the side effects of the teething troubles of the administrative reforms was the discontinuation of the Schematismus series for several years. Between 1849 and 1868 there are only eight editions kept at the ÖNB – none from 1849 to 1853.

Reasons for these gaps in the Schematismus series might be found in the ever strained financial situation of the state in this era[25] or in grave logistic hinderances, which initially complicated the lives of the officers that were sent to the provinces and might have prohibited the timely retrieval of reliable statistics for yearly publications.[26] For the 1860ies – with the February Patent of 1861 and the Compromise of 1867 a constitutionally similarly agitated decade like the 1850ies – there exist even fewer editions of the Schematismus. It is only in the later phase of the Habsburg monarchy that the the Schematismus returns to becoming an annual custom. The time series from 1876 through 1918 marks the longest continuous appearance of the Schematismus without intermission. These later years of the Schematismus series are characterized by consolidated content structure and print layout and will serve as an entry point for the Schematismus Project.

By taking the year 1900 as an example we can look at the overall contents of a typical edition for the later years of the series. The 1900 Schematismus is 1,400 pages long and is structured into six distinct parts. Namely, these are 1,070 pages of tabular and short-prose data, forming the main body of the book, a comparatively short 8-page table of contents and an extensive 259-page long name register, listing all personal names mentioned in the data. The remaining 50 pages labeled

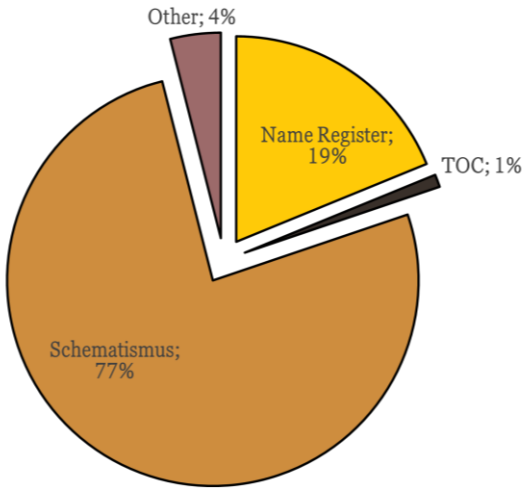


Figure 3: Metrics of the 1900-Schematismus.

[25] BRANDT Harm-Hinrich, Der österreichische Neoabsolutismus. Staatsfinanzen und Politik; 1848 - 1860. Zugl.: München, Univ., 09 - Fachbereich Geschichts- u. Kunstwiss., Habil.-Schr., 1974/75 (=Schriftenreihe der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften 15). Göttingen: Vandenhoeck und Ruprecht 1878, pp. 277–80.

[26] DEAK, Forging a multinational state, 2015, pp. 118–31.

"Other" in Figure 3 contain registers of titles and abbreviations and commercial advertisements.

2.2 Evolution of the Schematismus Layout

Every edition of the Schematismus series tries to depict the offices of the court and state in a structured way, giving insight into the institutional organization. Part of the organizational information, especially regarding hierarchical structures and departmental boundaries, is transported via the layout of the printed pages. There is a clear functional separation between headings, which always give the name of the office or other institutional entity and the left-aligned text below, which lists the names and further information of the officials occupying the different posts.

As mentioned earlier, reliably differentiating between different types of headings and left-aligned paragraphs of text will be critical for automated information extraction. The next few lines are therefore dedicated to a short overview of the evolution of the print layout of the Schematismus.

A periodical publication that existed for over 200 years was of course subject to major changes regarding its visual appearance, and not just because of different editors. While changing rulers, changing forms of government, the extent and structure of state administration determined the changes in content, industrialization of production technologies and different fonts changed the scope and appearance of the printed work.

The excerpts shown in Figure 4 illustrate the transition of layout and typesetting, which the Schematismus underwent during its publication. The early editions up to 1774 are printed on very slender pages, leaving space for just a single narrow column of text. The crowded layout leaves little room for stylistic differentiation between textual elements. The typographical differences between headings and paragraphs are very nuanced and semantic text understanding is important in order to structure the contents correctly. From 1775 onwards the printers of the Schematismus used a more modern aspect ratio of the pages. The screenshot from the 1800 Schematismus is representative for this phase. It shows headings that are centered and thus better separated from the left-aligned paragraph below. This layout, which still mainly uses bulk text in a single column is maintained but evolved during the first half of the 19th century. As the snippet from 1848 shows, bold scripts and brackets are introduced to further diversify heading styles. By that time other stylistic elements like curly brackets were already a common feature and two- and three-

column sections occur more and more frequently. The post-revolutionary modernization efforts are also reflected in the abandonment of the Fraktur script in the Schematismus. The printers used antiqua glyphs and the bulk of the officers were now listed in columns. First a two-column layout was used but by 1868 the three-column standard layout was established, which should be maintained until the end of the monarchy in 1918.

Except for the bold surnames, the extracts from the later editions of 1868 and 1918 show no major differences in terms of structure or layout, which suggests a consolidation of the entire printed work in its final phase.

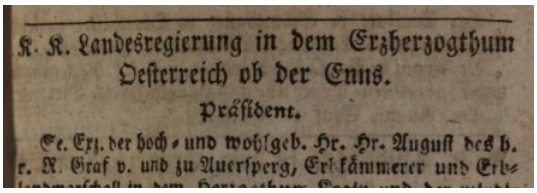
What will become clear in the later chapters of this thesis is that automated layout segmentation cannot be arbitrarily applied to any given page of written text. Good algorithms are normally fine-tuned to very specific tasks – for example a certain layout type. The consolidated layout of the later Schematismus pages serves therefore as an entry point for the Schematismus Project. The first project phase will be dedicated to extracting the years from 1876 to 1918 and the further considerations of this thesis will relate to this specific layout.

This concludes the initial overview of the Schematismus as a source for historical research. From here, there are two focus areas that need to be examined: The next chapter will address the type of data that is stored in the Schematismus – i.e. mass data on named individuals – and how the historical research specialization called prosopography deals with such data. The investigation will include a theoretical foundation and concrete examples. In the succeeding chapter the data extraction pipeline of the Schematismus Project will be discussed as a second focus area.

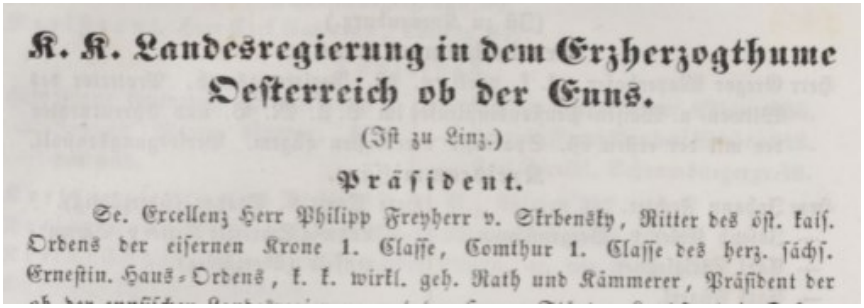
1702:


Käyſerl. Ob. De. geheime
Hof-Cantzley iſt d. Z. in dem
Käyſ. Hof-Spitall.
Ern Johann Georg von Buol/
Tyroliſ. Rittermäſſiger [illegible]mann/
der

1800:


K. K. Landesregierung in dem Erzherzogthum
Oesterreich ob der Enns.
Präsident.
Se. Exz. der hoch- und wohlgeb. Hr. Hr. August des h.
r. R. Graf v. und zu Auersperg, Erb-kämmerer und Erb-

1848:


K. K. Landesregierung in dem Erzherzogthume
Oesterreich ob der Enns.
(Ist zu Linz.)
Präsident.
Se. Excellenz Herr Philipp Freyherr v. Skrbensky, Ritter des öst. kais.
Ordens der eisernen Krone 1. Classe, Comthur 1. Classe des herz. sächs.
Ernestin. Haus-Ordens, k. k. wirkl. geh. Rath und Kämmerer, Präsident der

1868:


Erzherzogthum Oesterreich ob der Enns.
A. Landes-Vertretung.
Landeshauptmann.
Dominik Lebschy, R. d. L. O., Abt des Prämonstratenser-Stiftes Schlägl.
Stellvertreter.
Wiser Carl, Hof- u. Ger. Adv. in Linz.
Landes-Ausschuss.
C. Abgeordnete der Landes-Hauptstadt Linz.
Eigner Moriz, wie oben.
Mayer Ignaz, Schiffbaumeister in
7. Rohrbach.
(Neufelden, Lembach, Haslach, Aigen.)
Pöschl Joseph, jun., Lederfabrikant in Rohrbach [1]).

1918:


A. Landes-Vertretung.
Landes-Hauptmann.
Seine Exzellenz Johann Nep. Hauser, KZV. 1., FJO-Kt. m. St., Gh. R., päpstl. Haus-Prälat, Konsist. R., Rechsr. Abg.
Stellvertreter.
Jäger Ernst, KZV. 2., FJO-Kt., J. Dr., Adv u. Gem. R. in Linz.
Landes-Ausschuß.
Beisitzer.
Dirmeier Johann, Hausbes. in Ried (Innkr.).
Dürckheim-Montmartin Georg Friedrich, Gf., FJO-Off., Herrschaftsbes. in Hagenberg, Stellv. des L. Hptms. im L. Aussch.
Erdorf Josef, Gf., H. R. d. R.

Figure 4: Snippets from the main section from different Schematismus editions.[27]

[27] Digitized by the ÖNB and publicly accessible under: https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1702&page=90, https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1800&page=208, https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1848&page=570, https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1868&page=435, https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1918&page=1029, (last accessed: 06.01.2025).

3 Prosopography and Data Modeling

Automatically decoding the visual and semantic information of the Schematismus will yield access to data on many individuals that are known by name and share common properties like their affiliation with certain institutions of the Habsburg civil service. Historians have established and evolved a research field called prosopography that deals with such kinds of data. Nowadays prosopography relies on digital data processing. Any prosopographical research project must therefore evaluate how their digitized data should be stored and represented. The following pages are therefore dedicated to an introduction into the research field of prosopography and followed up by a brief overview of prosopography-specific data modeling standards.

3.1 Historical Prosopography

Definitions of prosopography are provided by Stone ("… the investigation of the common background characteristics of a group of actors in history…"[28]) or Verboven et al. ("… an attempt to bring together all relevant biographical data of groups of persons in a systematical and stereotypical way."[29]). At first glance, biographical research and genealogy both pursue quite similar goals. However, they can be clearly separated. Individual biography focuses on the collection of life data of a single person. Prosopographies, on the other hand, always deal with groups of individuals. The same is true for genealogical research. Here, the difference lies in the fact that genealogy, almost complementarily to biography, focuses on the connections within a family. In prosopography, kinship relations are not decisive. Instead, the studied group is defined by the research question. Verboven et al. summarize it by stating that biography and genealogy particularly focus on personal characteristics, whereas prosopography focuses more on non-personal characteristics. It is about what commonalities the examined groups have aside from the personal level.[30] From this perspective, one might think that prosopography is more closely related to sociology, or more specifically, to sociography. However, in sociography, the individual is sidelined, and society – or at least a part thereof – is examined in its entirety. The main distinction is

[28] STONE Lawrence, Prosopography. In: Daedalus 100, 1, 1971, pp. 46–79, here: p. 46.

[29] VERBOVEN Koenraad / CARLIER Myriam / DUMOLYN Jan, A Short Manual to the Art of Prosopography. In: Katharine S. B. KEATS-ROHAN (Ed.) Prosopography approaches and applications. A handbook (= Prosopographica et Genealogica 13). Oxford: Unit for Prosopographical Reasearch Linacre College Univ. of Oxford, 2007, pp. 35–70, here: p. 37.

[30] Ibid., pp. 39–40.

that prosopography always starts from a known – typically by name – group of individuals and focuses solely on them as its primary research subject. The social sciences require significantly larger and more complete datasets than are usually available to prosopography.[31]

The domain of prosopography, therefore, occupies an expansive territory between biography and sociography. Methodologically and conceptually, it tends to lean towards one pole or the other, depending on research context. When smaller, mostly elite groups are examined, comprehensive biographical data is typically available about each individual, allowing the research to also draw on qualitative methods. A study is considered prosopography, if it transcends the individual level and accesses biographical information only in aggregated form.[32] The study of less elite social groups usually involves significantly larger populations, with considerably lower informational density for the individual. Here, the use of quantitative and statistical methods, as is common in the social sciences, is widespread. Yet again, it is not only the thematic focus that determines the method, but also the time period and region being studied. The source density in the pre-modern era is much lower, which makes the use of quantitative methods more difficult.

The predecessors of prosopography can already be identified in the late 19th century. Not only did bureaucratization with its tendency towards documentation and archiving become fully established before and during the long 19th century,[33] but the urge to create collections was also omnipresent elsewhere. Specifically, national movements sought to assert their significance through the creation of national biographies.[34] The historiographical study of antiquity also took the Zeitgeist as an opportunity to initiate large encyclopedic projects. The bureaucratic and (historical) biographical data amassed throughout the century served as an incubator for the prosopographic approach. An examination of the data beyond the individual level ultimately imposed itself upon researchers. An example is provided by Verboven et al. in the context of Pauly's Realencyclopädie der classischen

[31] BAILLIE James, Alternative Database Structures for Prosopographical Research. In: International Journal of Humanities and Arts Computing 15, 1-2, 2021, pp. 117–132, here: p. 118.

[32] KEATS-ROHAN Katharine S. B., Biography, Identity and Names: Understanding the Pursuit of the Individual in Prosopography. In: Katharine S. B. KEATS-ROHAN (Ed.) Prosopography approaches and applications, pp. 139–182, here: pp. 140; 143; 151.

[33] OSTERHAMMEL Jürgen, Die Verwandlung der Welt. Eine Geschichte des 19. Jahrhunderts (=Historische Bibliothek der Gerda Henkel Stiftung). München: C.H. Beck [5] 2010, pp. 31-44; 878-882.

[34] CARTER Philip, What is National Biography For? Dictionaries and Digital History. In: Karen FOX (Ed.) True Biographies of Nations? The Cultural Journeys of Dictionaries of National Biography: ANU Press, 2019, pp. 57–78.

Altertumswissenschaft.[35] According to their description, "while collecting the information on senators and knights of the Roman Republic, the contours unexpectedly emerged of an aristocracy (nobilitas), which had built its power base and hierarchy behind the screens of the official institutions."[36] The motivation behind the prosopographic approach could not be illustrated any better. It is evident that certain historically effective structural characteristics transcend the individual level and only become tangible when one examines the actors based on their common properties and considers them as a group. The breakthrough as an important approach in modern historical research came in the first half of the 20th century. Stone[37] locates the beginning of professional prosopographic research in the 1920s and 1930s. During this time, the thematic focus opened significantly, and alongside ancient collections, the records of modern bureaucracies and other historical periods and disciplines came under scrutiny.[38]

3.2 Prosopographical Sources and Data Collection

Historical Prosopography needs to collect and aggregate personal data from source material. The availability of data and the effort associated with its collection and aggregation vary between research contexts and the historical epoch, but some challenges are shared between most prosopographical projects.

One of them is the problem of reliably identifying the individual. Prosopography always starts from the individual. While final statements are to be made about social groups, these statements are always derived from aggregated individual characteristics. The individual should remain identifiable in the basic data. As in everyday life, the most common identifier for a person in historical sources is their name. However, name similarity is usually a far less grave problem in everyday life than in prosopographic research. Prosopographic research must take into account that several people can have the same names, but also that the same person can be mentioned in the sources under different names.

Another common problem relates to a lack of uniformity and comparability of historic records. Ideally, a source document would include a person by name and the related

---

[35] The editions of the "Ur-Pauly" have been digitized and can be accessed online: https://de.wikisource.org/wiki/August_Friedrich_Pauly#Herausgeberschaft (last accessed: 01.02.2025)
[36] VERBOVEN / CARLIER et al., A Short Manual to the Art of Prosopography, 2007, p. 43.
[37] STONE, Prosopography, 1971.
[38] VERBOVEN / CARLIER et al., A Short Manual to the Art of Prosopography, 2007, p. 43.

description would use standardized wording so that properties like age, occupation, relations and other affiliations could be marked up and collected directly. However, most historical source texts are far from being standardized biographies. Even in the Early Modern period life paths and professions were far less standardized than nowadays and so was their textual description. Even within small geographical and social stratifications differing terms could describe the same concept. Thus, finding comparable properties is often connected with interpretative subsummations of different terms under a common concept.

The following lines will try to exemplify the discussed difficulties citing a publication, which was part of a large-scale research project examining the "Republic of Letters". The "Republic of Letters" was a network of scholarly communication that existed from the mid-15th century until around 1800, maintained through letters across political and institutional boundaries. Admission to and acceptance within this circle was solely based on the ability to assert oneself intellectually. Consequently, its members came from all social classes and religious denominations, including men and women, aristocrats and craftsmen. Under such circumstances, collecting common personal features proves more difficult than it might seem and be it just about a subset of this population. Additionally, the long observation period and the trans- and even intercontinental distribution of individual members introduce further variations that complicate the aggregation and comparability of information between individuals. For members of the administration and the legal system, one quickly finds that a multitude of political frameworks and bureaucratic systems make standardizations challenging even within a limited timeframe. Different hierarchies, different designations for similar offices, or even tasks that are not comparable make structured data collection difficult. In the context of the Republic of Letters Hotson, Wallnig et al. recommend focusing on members of universities and religious orders for an initial attempt at modeling, as these institutions have established and maintained relatively stable structures throughout the entire period under consideration and additionally have carefully documented their members and their activities.[39]

[39] HOTSON Howard / WALLNIG Thomas / JENSEN Mikkel Munthe et al., Prosopographies of the Republic of Letters. In: Howard HOTSON / Thomas WALLNIG (Eds.) Reassembling the republic of letters in the digital age. Standards, Systems, Scholarship. Göttingen: Universitätsverlag Göttingen, 2019, pp. 371–398, here: p. 372.

3.3 Types of Digital Prosopographies

The prosopographical research method is often described as a two-part problem of data collection and data evaluation. The more traditional point of view sees them as sequential steps of a process. The first step involves the collection and systematic compilation of all available biographical information. The group of individuals to be studied is identified, a name index created, and the information made accessible through a database. The second step encompasses the examination of the collected data through one (or more) structured 'data-capturing' questionnaires to filter the information needed from each individual to answer a specific research question.[40]

Especially in the last few decades, developments in information technologies have created new possibilities of data storing and data analysis and have widened the scope of prosopographical research. Today prosopographical projects can take many forms depending on their technological implementation. The shape and fragmentation of the source material is still a decisive factor but increasingly so are user experience and analytical capabilities. Thus, the field of prosopography has become less uniform, and the picture of the two-step process loses its universal applicability.

Baillie gives some guidance on how modern prosopographical projects could be meaningfully categorized. He distinguishes between prosopographical indexes and secondary prosopographies. Prosopographical indexes are strongly source-oriented and provide a database that facilitates future work and allows for references to the original documents. They originate from research environments where information is scattered over many fragmented source documents, and their added value lies in providing a central searchable store to channel the effort that has flown into finding and collating the sources. Baillie describes them as a "reference tool in order to find relevant pieces of source material or determine a particular figure's identity."[41]

Because of their close relation to the original source material prosopographical indexes can easily represent conflicting source statements. If they do so, however, they present those conflicting statements on an even footing.[42] What prosopographical indexes are less

---

[40] KEATS-ROHAN, Biography, Identity and Names: Understanding the Pursuit of the Individual in Prosopography, 2007, pp. 146–147.

[41] BAILLIE James, Alternative Database Structures for Prosopographical Research. In: International Journal of Humanities and Arts Computing 15, 1-2, 2021, pp. 117–132, here: p. 123.

[42] Ibid., p. 128.

ideally positioned to do is to represent interpretative or analytical information, i.e. a scholarly assessment of the credibility of conflicting statements. This is the domain of secondary or analytical prosopographies. Secondary prosopographies, as envisioned by Baillie, use data models that allow them to represent causal relations.[43] This makes them suited for offering consistent readings of causal networks that can be traversed over a multitude of paths and analyzed through their network connections. Secondary prosopographies do not neglect or hide their connection to the original source but their main purpose is to facilitate analysis on the prosopographical database itself rather than offering a structured guide to the original texts.

Especially prosopographies that are based on rich data sources benefit from the analytical ease of modern technological solutions and readily try to implement them. Graph databases in particular are ideally positioned to address many of the data modeling challenges of analytical prosopographies. The next chapter is therefore dedicated to give an introduction to prosopographical data modeling with a special consideration of graph databases.

3.4 Data Modeling for Digital Prosopographies

Like most present-day prosopographical projects, the Schematismus project derives value from creating a cross-linked database. This database enables users to trace individuals' institutional affiliations and visualize the civil service's institutional structure. We conceptualize a database as the repository where digital data is stored in a structured format. Creating a database requires some initial design considerations, collectively known as data modeling.[44] This process entails abstracting and standardizing perceived information into discrete entities and relations, each characterized by specific attributes.[45] The database's ultimate value lies in providing streamlined options for searching, aggregating, and displaying information—capabilities that justify the considerable effort invested in its development.

---

[43] Ibid., pp. 118–19.

[44] KLINKE Harald, Datenbanken. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 109–127, here: pp. 109–112.

[45] JANNIDIS Fotis, Grundlagen der Datenmodellierung. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities, pp. 99–108, here: pp. 102–103.

Data modeling constitutes a core discipline within the Digital Humanities, with its significance expanding considerably alongside the emergence of semantic web technologies and AI. Beyond these advanced applications, the fundamental importance of data modeling is evident at even the most basic level. Essentially, it is about making the humanist's knowledge and assumptions about historical events or artifacts tangible for computer-based processing.[46] Scholars in the field note that "only when data is imbued with meaning (semantics) can it be referred to as 'information'" thus highlighting the transformative process at the core of data modeling.[47] A data model is essentially an attempt to describe an original – whether it be an object, an action, or a concept – based on characteristic features, with the resulting advantage of being able to perform complex computer-based operations on this model subsequently.[48] Modeling thus also refers to the ability to simulate observable effects through variables and inputs.[49] With regards to prosopography, data modelling encompasses, therefore, considerations on the abstracted digital representation of personal relations extracted from historical sources by historical researchers.

3.4.1 The Factoid Model

Since the start of professional prosopography Britain has been a center for prosopographical research. Since the 1990ies the King's College in London created the Prosopography of the Byzantine Empire (PBE), the Prosopography of Anglo Saxon England (PASE) and many more, all of which are among the most prominent projects of digital prosopography. The Factoid Model was a product of these large-scale endeavors and established itself as a widespread standard for prosopographical data modeling.

A factoid is an assertion made by a scholar that a certain passage of a source text describes a certain relationship. The prosopographically relevant elements – people, objects, actions, etc. – are extracted and linked back to the factoid. The name "factoid" refers to the idea that a source extraction taken out of its overall and interpretative context should not be considered as a fact. For example, contradictory information about the same person can appear in different sources. As the standard model for prosopographical indexes, the strengths of this approach lie in making all these various statements visible and presenting

---

[46] Ibid., pp. 107–8.
[47] KLINKE, Datenbanken, 2017, p. 109.
[48] JANNIDIS, Grundlagen der Datenmodellierung, 2017, p. 100.
[49] OLDMAN Dominic / DOERR Martin / GRADMANN Stefan, Zen and the Art of Linked Data. In: Susan SCHREIBMAN / Raymond George SIEMENS / John UNSWORTH (Eds.) A new companion to digital humanities (= Blackwell companions to literature and culture 93). Chichester, West Sussex – Malden, MA et al.: Wiley Blackwell, 2016, pp. 251–273, here: p. 257.

them in a close-to-source manner, thereby making them accessible for interpretation.[50] The structuring of the data according to this approach is done in a "source-driven" manner.[51] Secondary interpretative influences should thus be kept to a minimum.[52]

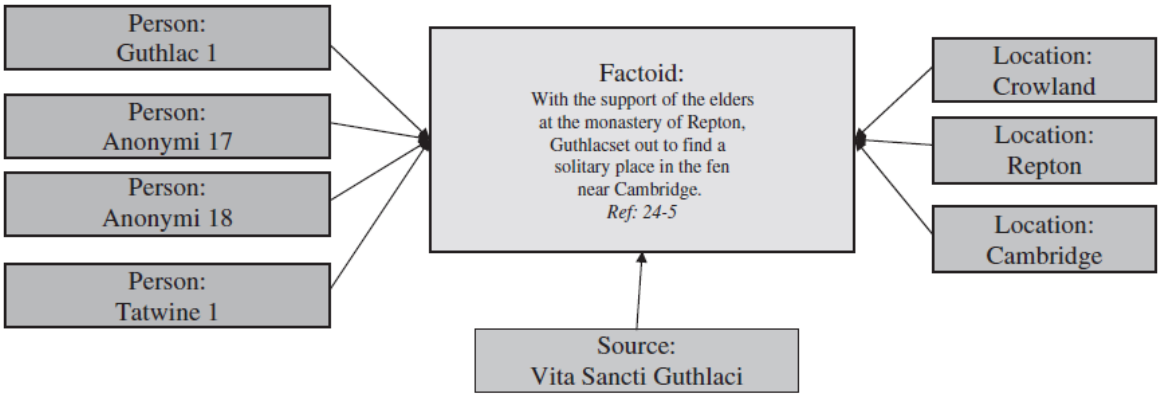


Figure 5: An example of a factoid and a set of information that is typically extracted from it.[53]

Factoid data is stored in an entity-attribute-relationship model and commonly implemented in a relational database.[54] Relational databases are established standard for many applications in the Digital Humanities.[55] They consist of various tables – one for each entity type. Using the example shown in Figure 5, the relational database for this project would at least contain tables for individuals, locations, factoids and sources. Each element shown in Figure 5 would occupy an entry in the respective table and be assigned with a unique key. The relations between different entities would be constructed by referencing those keys.[56]

The first projects using the Factoid Model date back to a time when natural language processing was still significantly more difficult, and projects in the field of computer-assisted historical research often dealt with transferring already highly structured data.[57] Nevertheless, storing data in a database enabled much more efficient exploration than

---

[50] BRADLEY J., Texts into Databases: The Evolving Field of New-style Prosopography. In: Literary and Linguistic Computing 20, Suppl 1, 2005, pp. 3–24, here: pp. 8–9.

[51] PASIN Michele / BRADLEY John, Factoid-based prosopography and computer ontologies: towards an integrated approach. In: Digital Scholarship in the Humanities 30, 1, 2015, pp. 86–97, here: p. 89.

[52] BAILLIE, Alternative Database Structures for Prosopographical Research, 2021, p. 119.

[53] BRADLEY, Texts into Databases: The Evolving Field of New-style Prosopography, 2005, p. 9.

[54] PASIN / BRADLEY, Factoid-based prosopography and computer ontologies: towards an integrated approach, 2015, pp. 86–89.

[55] JANNIDIS, Grundlagen der Datenmodellierung, 2017, p. 101.

[56] KLINKE, Datenbanken, 2017, pp. 112–122.

[57] BRADLEY, Texts into Databases: The Evolving Field of New-style Prosopography, 2005, p. 4.

previously possible. For example, it allowed searching based on specific attributes, grouping individuals according to those attributes, or creating complex queries through combinations.[58]

The Factoid Model was an innovative approach at the time of its creation. Its sustained popularity[59] for modelling prosopographical data warrants that the Factoid Model is still well suited for many purposes and sufficiently user friendly. As mentioned in the previous chapter, the Factoid Model is preferably used for prosopographical indexes. It is well-suited for gathering and structuring dispersed source texts.

3.4.2 Data in the Semantic Web and Ontologies

In recent years the semantic interlinking of data in connection with semantic web technologies and Linked Open Data (LOD) has gained importance. The Semantic Web is intended to be an extension or evolution of the Internet. The published data should be semantically enriched and able to be linked to one another, which, in its best vision, leads to globally interconnected knowledge and linked data.[60] These innovations build on causal data networks and represent knowledge in the form of causal triples. The Resource Description Framework (RDF)[61] is the standard meta-model for these triple representations in the Semantic Web. A triple consists of a subject, predicate, and object – a structure that enables relational connections between information units. For instance, "Muhammad Ali" (subject) and "Rumble in the Jungle" (object) become meaningfully linked through the predicate "won".

< Muhammad Ali >	< won >	< Rumble in the Jungle >
Subject	*Predicate*	*Object*

Subject and object can be connected through further predicates with other information units, such as "person", "boxer", "George Foreman", "Kinshasa", or 1974.

Clearly, such a form of knowledge modeling requires conventions and standards. Predicates cannot be arbitrary but must be selected from abstract classes. Only in this way can a reduction of the real-world relation to a model-like schema be implemented. For example: <Robert Musil> - <wrote> - <The Man Without Qualities>. The predicate "wrote"

---

[58] PASIN / BRADLEY, Factoid-based prosopography and computer ontologies: towards an integrated approach, 2015, p. 88.

[59] BAILLIE, Alternative Database Structures for Prosopographical Research, 2021, p. 119.

[60] OLDMAN / DOERR et al., Zen and the Art of Linked Data, 2016, p. 252.

[61] CYGANIAK Richard / WOOD David et al., RDF 1.1 Concepts and Abstract Syntax 2014. Online reference: www.w3.org/TR/2014/REC-rdf11-concepts-20140225/, (last accessed: 01.02.2025).

expresses that the subject created the object. “Writing” can therefore be subsumed more generally under the concept of “creating”. This step is important for the model-like abstraction. The abstracted triple can then be specified more precisely if needed through adjacent nodes (writer, novel) or added properties (created – type: writing).

This need for standardized abstraction and normalization processes has led to the development of ontologies. Ontologies are schematic representations of knowledge that enable the integration of heterogeneous data sources and logical conclusions.[62] In addition to other classifications, a distinction can be made between upper ontologies and domain ontologies. Upper ontologies have the highest level of abstraction and aim to formalize all conceivable relationships. They pursue a universal claim. Just as everyday language is often insufficient for communication in scholarly research disciplines, the complex causal relationships in these fields require tailored domain ontologies.[63] Originating from the museum and collection sector, CIDOC-CRM has developed into the standard for most cultural heritage applications.[64]

Data representation as causal triples holds several advantages for prosopographical projects. Triples are well suited to describe and analyze prosopographical networks. Relations between individuals and institutions can be abstracted to triples using predicates like <is related to>, <is owner of>, <is member of> etc. Using the right techniques, such networks can be efficiently represented and analyzed. Additionally, their causal structure facilitates data representation as a narrative. This, and their connectivity to LOD resources can increase user experience and user value. For example, names of known individuals can be connected to their entry in the GND[65] and expanded through this information store. Triple data representations are especially suitable for secondary prosopographies as defined in the previous chapter 3.3.

Even established prosopographical data models have seen the advantages of implementing structures that can be connected to the Semantic Web. The creators of the Factoid

---

[62] REHBEIN Malte, Ontologien. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities, pp. 162–176, here: p. 162.

[63] Ibid., pp. 165–66.

[64] CROFTS Nicholas / DOERR Martin et al., The CIDOC Conceptual Reference Model: A standard for communicating cultural contents. Online reference: cidoc.ics.forth.gr/docs/martin_a_2003_comm_cul_cont.htm, (last accessed: 01.02.2025). The current standards and recent developments can be found online: https://cidoc-crm.org/ (last accessed: 02.05.2025)

[65] The GND is a publicly available collection of cultural and research authority data curated by the German National Library: https://gnd.network/Webs/gnd/DE/Home/home_node.html (last accessed: 02.02.2025)

Model have conceptually restructured their model architecture for formalization in CIDOC-CRM.[66] The creation of a factoid is in this context conceived as a Document Interpretation Act, which is conducted by an agent, based on a document, and in which a context of action and a temporal entity are extracted. This schema is mapped to selected classes of the CIDOC-CRM ontology.

The creation of the factoid by a historical scholar is perceived as "document interpretation act". The document interpretation act is the central entity from which three triples emanate:

- <factoid: Document Interpretation Act><has_author><factoid: Agent>
- <factoid: Document Interpretation Act><has_source><factoid: Document>
- <factoid: Document Interpretation Act><has_interpretation><factoid: Temporal Entity>

In the previous chapter the factoid was introduced as a short extract from a source text, which describes a certain historical event. In the new model this text extract is assigned to the <factoid: Temporal Entity>-entity. It is modeled as the result of an interpretation process. The same process has a source document <factoid: Document> and an interpreter <factoid: Agent>. It may seem cumbersome and chaotic to re-map orderly compiled tables to countless triple representations, but at the same time the new model's intrinsic causal structure, which facilitates network analysis and user experience, becomes immediately recognizable.

In addition to the re-modeling of the standard Factoid Model to semantic triples, Figure 6 shows how the elements of the triple-representations are mapped to standard categories of the CIDOC-CRM ontology. CIDOC-CRM uses entities and properties as functional categories. When applied to RDF-triples, subject and object are entities, while the predicate is a property. From the variety of generalizations that the CIDOC-CRM-standard offers, only a few are selected as suitable generalizations for the relations of the factoid representation. So, for example, the <factoid: Agent> is an "E39 Actor"-entity, while the <factoid: Document> is an "E73 Information Object"-entity. Utilized properties are for example "P14 carried out by" or "P141 assigned".

---

[66] PASIN / BRADLEY, Factoid-based prosopography and computer ontologies: towards an integrated approach, 2015, pp. 91–93.

More information on the mapping of the Factoid Model to the CIDOC-CRM ontology can be found in the referenced literature. This chapter of the thesis will in the following briefly investigate alternative database structures, which are more efficient at storing and analyzing network structures than the traditional relational databases.

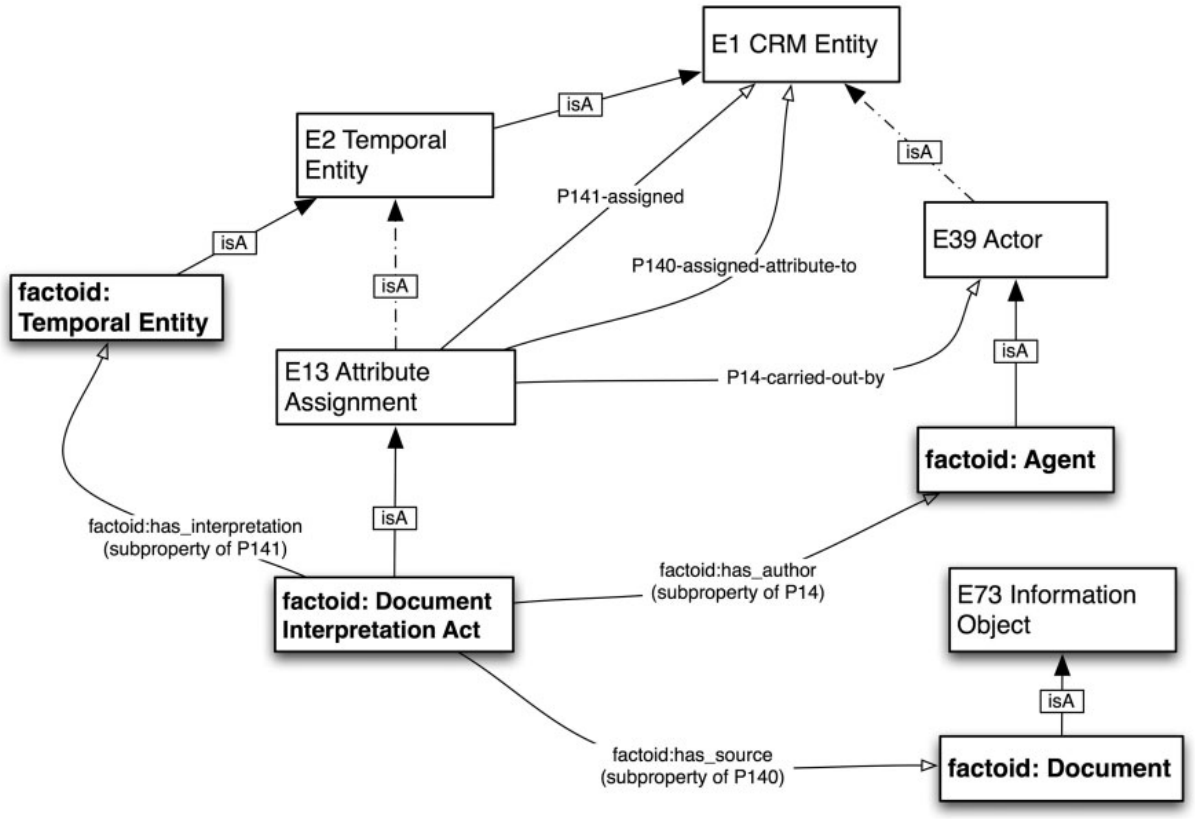


Figure 6: Conceptual revision of the factoid architecture to facilitate mapping to CIDOC-CRM.[67]

3.4.3 Graph Databases

Changed demands regarding data representation have also implications on the database technologies used. Relational databases are not very efficient at representing and analyzing large networks. To be able to access individual entities, relational databases need to load the whole respective table into memory. The growing importance of semantic web technologies and Linked Open Data (LOD) has therefore driven the evolution of graph databases.

Grounded in mathematical graph theory, these databases excel particularly in network analysis. The underlying theory facilitates calculations concerning neighborhoods and

[67] Ibid., p. 93.

pathways within networks structured as nodes and edges. Typical tasks include finding the shortest paths or centrality measures for individual nodes. If the edges can only be traversed in one direction, they are referred to as directed graphs.[68] Graph databases bring several advantages, especially in querying. This is because queries are made based on a node with respect to its neighborhood. This allows environments to be visualized or paths to be followed further. Furthermore, the queries are extremely efficient because only the respective objects with their neighbors need to be loaded, rather than the entire database.[69]

But graph databases are not only ideally positioned to address the analytical needs of prosopographical research. Directed graphs offer an intuitive way to store RDF-triples, when modeling subject and object as nodes and the predicate as directed edge.[70] Modern graph database systems like Neo4j use a data model that is capable of further enriching nodes and edges with attributes.[71]

The many opportunities that graph databases offer have led to several suggestions to use them in the context of prosopographical data representations. Akoka et al., for example, have proposed an Uncertainty Knowledge Graph solution, in which contradictory statements are juxtaposed, and each triple can be assigned probability values. The probability value should reflect the assessment of domain experts and thus depict the negotiation process of scientific practice.[72] This model specifically addresses a frequently discussed, central problem of the Factoid Model, since it cannot resolve uncertainty about the truthfulness of the source and its contents.

Another shortcoming of the Factoid Model, which is addressed in the so-called person-prosopon model by Baillie, are implicit properties that allow researchers to add their scholarly expertise.[73] Baillie's model is, however, not intended and also not suited to replace the Factoid Model in its realm of prosopographcal indexes. The person-prosopon model is a model for secondary prosopographies, as it cannot integrate contradictory sources

---

[68] JANNIDIS Fotis, Netzwerke. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities, pp. 147–161, here: pp. 147–153.

[69] KLINKE, Datenbanken, 2017, p. 126.

[70] BAILLIE, Alternative Database Structures for Prosopographical Research, 2021, p. 120-121.

[71] REHBEIN, Ontologien, 2017, p. 172.

[72] AKOKA Jacky / COMYN-WATTIAU Isabelle / LAMASSÉ Stéphane et al., Contribution of Conceptual Modeling to Enhancing Historians' Intuition -Application to Prosopography. In: International Conference on Conceptual Modeling (ER) Vienna, Austria, 2020, here: p. 53.

[73] BAILLIE, Alternative Database Structures for Prosopographical Research, 2021, p. 127.

sufficiently but rather models a consistent-reading, which is based on the historically most viable interpretation.[74]

Another graph-based approach is being developed as part of the RELEVEN project by Andrews et al. at the University of Vienna.[75] The RELEVEN project is an ERC-funded endeavor to reconstruct the Christian world of the "short eleventh century" between 1033 and 1095. Besides historical challenges such as the multiplicity of languages, cultural contexts, and types of evidence, the authors also strive to implement a new data model that focuses on "scholarly interpretation and source provenance," with the aim of overcoming claims that data-driven historical research is too positivistic.[76] By integrating interpretation and source provenance – also of conflicting sources – this approach tries to unify the different virtues of some of the aforementioned methods.

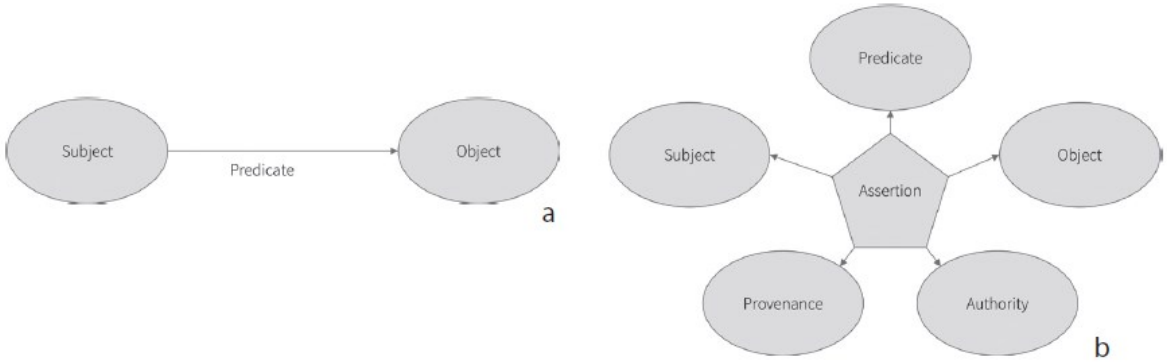


Figure 7 Schematic expansion of the classic RDF triple to the five-part STAR model.[77]

The model named STAR extends the classic graph triple to a five-part formation consisting of subject, predicate, object, source, and authority. The STAR model can also be used to expand existing graph triples by adding a source and an authority.[78] Andrews et al. use the CIDOC-CRM ontology as a template for their STAR model and overlay it with a

[74] Ibid., p. 124.
[75] https://releven.univie.ac.at/ (last accessed: 11.05.2025)
[76] ANDREWS Tara / RÓZSA Márton / PRAJDA Katalin et al., Re-Evaluating the Eleventh Century through Linked Events and Entities. In: Historical Studies on Central Europe 4, 1, 2024, pp. 217–245, here: pp. 217–219.
[77] Ibid., p. 220.
[78] ANDREWS Tara, The STructured Assertion Record (STAR) Model for Event-based Representation of Historical Information. Mainz 2023, p. 1. Online reference: graphentechnologien.hypotheses.org/files/2023/05/GrapHNR-2023-32-Andrews-STAR.pdf, (last accessed: 08.02.2025).

custom ontology to represent the perspectives of different authors (in the model, authorities). These perspectives consist of statements that can either be accepted or rejected.[79]

The advantages of triple representations and graph databases have convinced the project team of the Schematismus Project to use them for the creation of their database. Since the information in the Schematismus is easy to find but much more difficult to interpret, the main intention of the project team is to create a prosopograraphical database that facilitates network analysis and at the same time preserves the reference to the original source. Data modeling for the Schematismus Project therefore demands the creation of effective triples from the extracted data. Before data extraction and data modeling of the Schematismus Project are explored in more detail, the next pages present important, recent prosopographical projects relating to the Habsburg court and state.

3.5 Digital Prosopographies of the Habsburg Empire

As the previously presented prosopographical projects indicate, digital prosopography – mostly in the form of prosopographical indexes – traditionally has strongholds in ancient and medieval history. From the early modern period onward, source information density increases dramatically. This shifts the research interest from locating sparse source documents to efficiently analyzing the dense data available. The following section will briefly introduce three projects of digital prosopographies of the Habsburg monarchy from early modern and modern sources.

3.5.1 The Austrian Prosopographical Information System

The Austrian Prosopographical Information System (APIS) is a virtual research environment[80] developed to transform the Austrian Biographical Dictionary (Österreichisches Biographisches Lexikon, ÖBL) into a prosopographical database.[81] Although this sounds like a very similar task to the digitization of the Schematismus, there are some notable differences. For one, the APIS project worked with texts that were already digitized and

[79] Ibid., p. 2.

[80] KAISER Maximilian / ROMBERG Marion / SCHLÖGL Matthias et al., Von APIS zu VieCPro. In: Helmuth ALBRECHT / Michael FARRENKOPF / Helmut MAIER et al. (Eds.) Historische Biographik und kritische Prosopographie als Instrumente in den Geschichtswissenschaften (= Veröffentlichungen Aus Dem Deutschen Bergbau-Museum Bochum 257). Berlin – Boston: De Gruyter Oldenbourg, 2023, pp. 117–140, here: p. 117.

[81] FEIGL Roland / GRUBER Christine, Das Projekt APIS und der ÖBL-Textkorpus in seiner digitalen Transformation. Herausforderungen für ein traditionelles biographisches Lexikon. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS) pp. 9–18, here: pp. 10–12.

secondly, while the Schematismus is a highly structured collection of defined data on a clearly circumscribed group of people, the criteria of who was added to the ÖBL and with which information was much less standardized. Generally speaking, there were less people recorded in the ÖBL, but each person with a much higher information density.[82]

Considering this, the challenge regarding data collection for the APIS project lay in the extraction of structured data from longer text passages. The system that was tailored to the task by the project team consisted of a procedural step to resolve abbreviations, a named entity recognizer, an entity linker and several components for disambiguation.[83] Named entity recognition struggled with problems that are typical for historical texts. For example, there are many often unorthodox abbreviations and many proper names to describe Institutions or professions are no longer used today and thus not familiar to natural language processing (NLP) algorithms.[84]

At the core of the APIS project is a relational database that builds on a custom APIS data model. As shown in Figure 8, the model knows five different entities (Person, Institution, Place, Work, Event) that can be linked with each other via relations – the white fields in the scheme.[85] This still leaves a lot of freedom to fill the properties of the entities and relations. For the ÖBL data, the APIS makes use of controlled Vocabularies and LOD-Ressources like GND to fill the properties in a structured manner.[86] The APIS data model can be exported into CIDOC CRM format to be connectable to LOD-standards.

---

[82] BERNÁD Ágoston Zénó / LEJTOVICZ Katalin, Korpusanalyse und digitale Quellenkritik – Die Vermessung des Österreichischen Biographischen Lexikons. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS) pp. 107–158.

[83] LEJTOVICZ Katalin / SCHLÖGL Matthias, Die (semi-)automatische Verarbeitung biographischer Artikel im APIS-Framework. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS) pp. 49–82, here: pp. 54–64.

[84] SCHLÖGL Matthias / LEJTOVICZ Katalin, Die APIS-(Web-)Applikation, das Datenmodell und System. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS). Vom gedruckten Textkorpus zur Webapplikation für die Forschung. Wien – Hamburg: new academic press, 2020, pp. 31–48, here: p. 31.

[85] Ibid., pp. 35–37.

[86] KAISER / ROMBERG et al., Von APIS zu VieCPro, 2023, pp. 120–121.

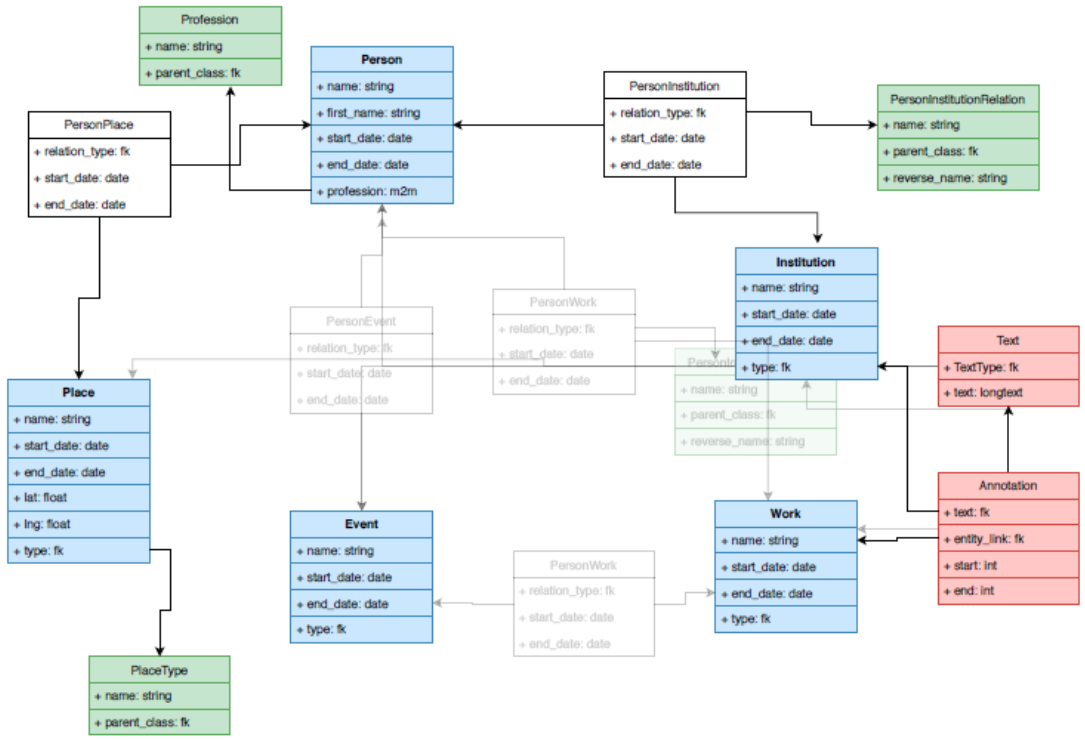


Figure 8: Simplified scheme of the APIS data model.[87]

The APIS data model proved to fit well also for other prosopographical data and the APIS platform has hence been used by a number of other research projects to make their database digitally accessible.

3.5.2 Viennese Court Prosopography Portal

Among these other projects is the ”Viennese Court Prosopography Portal” (VieCPro), which also demands our attention as a recent prosopographical project in the field of Habsburg history. The aim of VieCPro lies in creating a comprehensive reference portal for the Viennese court from the middle of the 17th century until the beginning of the 19th century.[88] Both projects APIS and VieCPro build largely on data that was digitized in earlier projects, but for both projects it was essential to enrich the digitized data via NLP or manually and build LOD linkages.[89]

[87] SCHLÖGL / LEJTOVICZ, Die APIS-(Web-)Applikation, das Datenmodell und System, 2020.
[88] KAISER / ROMBERG et al., Von APIS zu VieCPro, 2023, p. 118.
[89] Ibid., pp. 118–19.

The data of VieCPro had previously been collected and digitally indexed but not digitally published during two completed research projects.[90] A big part of the data from the 18th century was transcribed from the "kaiserlicher Hof- und Ehrenkalender", which shares many similarities with the Schematismus. For the years 1711-1765 the data holds the prosopography of 4000 officers of the Habsburg court.[91] A dataset that should for the biggest part match the respective data from the Schematismus of these years. In later stages of the Schematismus Project, when also the earlier editions from the 18th century will be digitized, the VieCPro database could therefore be used as valuable point of reference.

3.5.3 Minutes of the Council of the Ministers of Austria

The "Ministerratsprotokolle Österreichs und der Österreichisch-Ungarischen Monarchie 1848–1918" (MRP) are an ongoing project that aims at creating critical editions of the minutes of the council of the ministers of Austria and of the Austro-Hungarian Monarchy in a hybrid form of print and digital. The print versions have been edited since 1970. Since 2018 the Austrian Academy of Sciences edits digital critical editions of the minutes in parallel to the print version.[92]

The digital version is implemented as TEI-XML[93] and is not a prosopographical project per se. The transition to a digital editing practice, however, encompasses enriching the textual representation with linked data.[94] The minutes of the council of ministers abundantly contain names of individuals and institutions of the state as well as places and other named entities. Some of these entities can be linked with databases of normed data – for example the GND – but for many instances no such data is available. To be able to enrich these instances with auxiliary data, the project team used the APIS platform and the APIS data

---

[90] ROMBERG Marion / KAISER Maximilian, The Viennese Court. A Prosopographical Portal. Eine Projektvorschau auf ein Referenz- und Nachschlageportal zum Wiener Hof von Leopold I. bis Franz II. (I.). In: Akademie der Wissenschaft zu Göttingen (Ed.) Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen. Neue Folge: Stadt und Hof; 9: Göttingen Academy of Sciences, 2020, pp. 43–52, here: p. 47.

[91] KUBISKA-SCHARL Irene / PÖLZL Michael, Adelige und bürgerliche Karrierewege bei Hof. Eine Prosopographie des Wiener Hofpersonals 1711–1806. In: Akademie der Wissenschaft zu Göttingen (Ed.) Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen. Neue Folge: Stadt und Hof; 3: Göttingen Academy of Sciences, 2014, pp. 76–86, here: pp. 76–79.

[92] The digital editions can be downloaded as data dump: KURZ Stephan, oeaw-ministerratsprotokolle/mp-edition-data: v. 1.5 including CMR calendar data 1872–1914. Wien 2024. Online reference: zenodo.org/records/11484662, (last accessed: 08.02.2025).

[93] SCHÖCH Christof, Ein digitales Textformat für die Literaturwissenschaft: Die Richtlinien der Text Encoding Initiative und ihr Einsatz bei Textkonstitution und Textanalyse. In: Romanische Studien 1, 4, 2016, pp. 325–364.

[94] KURZ Stephan / FISCHER-NEBMAIER Wladimir / KAMPKASPAR Dario et al., Die Edition der Ministerratsprotokolle 1848-1918 digital. Workflows, Möglichkeiten, Grenzen. In: digital humanities austria 2018. empowering researchers. Vienna: Austrian Academy of Sciences Press, pp. 83–90, here: pp. 84–88.

model to create a prosopographical database.[95] This database contains names and institutions of the civil service that are mentioned in the minutes. The project team tried to mirror the hierarchical institutional structure in the database. To that end, they manually transcribed information from the 1900 and 1917 editions of the Schematismus to excel tables.[96] As of today the named entities in the MRP link to this institutional tree. This organizational structure is static and the organizational tree from 1900/1917 might in fact not fit references from earlier years.

The MRP were added as an example to highlight how beneficial a continuous reconstruction of the organizational structure and personnel of the institutions of the Habsburg state could be for other research projects. Cross-Linking of the databases could immensely increase the depth and accuracy of auxiliary data available for information enrichment.

This brief review of recent prosopographical projects of the Habsburg court and state not only highlights how some of these projects might benefit from the database of the Schematismus Project but also shows that none of them has so far used knowledge graphs as database technology. The Schematismus Project will therefore be in many regards a valuable addition to the research landscape.

[95] The database is named "Modulare Prosopographische Registratur (MPR)" and can be accessed and queried online: https://mpr.acdh.oeaw.ac.at/ (last accessed: 02.02.2025)

[96] Stephan Kurz liberally provided insights into their processes and manually extracted excel files.

4 Data Extraction and Data Modeling of the Schematismus Project

Following the examination of the intricacies prosopographical data modeling topics and associated research projects of chapter 3, the focus will return to the data extraction from the Schematismus. Prior to chapter 3, it was observed that the style and dimensions of headings in the Schematismus serve as a reliable indicator of the intrinsic institutional structure and hierarchy. Furthermore, it was noted that these characteristics are subject to change over time. The alterations in layout and font outlined in section 2.2 exert limited influence on readability for humans. However, they constitute substantial impediments to attempts at automatic information extraction.

The following pages offer a comprehensive introduction to the technologies that underpin the Layout Model of the Schematismus Project. The concise section makes repeated use of technological terminology and references specific models from the domain of computer vision. The objective of this text is not to provide an exhaustive exposition of the technologies in question, but rather to offer the reader an overview of the technologies employed and points of reference that will facilitate further exploration of the topic.

4.1 The Technological Basis of Extraction

The information that should be decoded and automatically extracted from the scanned pages of the Schematismus can be attributed to the typography and layout of the page – i.e. geometric orientation, coherence of text blocks, font and special characters placed outside the text flow. Although the automated extraction of all these information components is based on optical characteristics and can therefore be assigned to the field of computer vision, a distinction must be made between at least two subtasks. Optical character recognition (OCR) is aimed at the digital reconstruction of the text contents. The technological basis is established and clearly predates current deep learning methods.[97] In comparison, the automated recognition of more complex layout components was, until recently, a much more difficult task. Established OCR algorithms can typically segment text lines and

[97] REHBEIN Malte, Digitalisierung. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 179–198, here: pp. 193–195.

columns. But the built-in algorithms don't render usable results on more complex layouts.[98] Especially since the advent of convolutional neural networks (CNNs), automatic image classification and segmentation have made progresses that were previously thought impossible.[99]

Artificial Neural Networks (ANN) are AI-models that utilize machine learning to adapt to complex and specific tasks. The CNN architecture – a special form of ANN – works with image data and has also proven to produce good results in document classification[100] and document segmentation tasks.[101] A further advantage of CNNs is that their hierarchical information layers allow transfer learning. This process – also known as fine-tuning – utilizes the fact that the deeper layers store learned representations of very general image features such as edges and only readjusts the higher layers, which represent task specific image features. The described process of fine-tuning is a supervised machine learning task, which is much more efficient than training a new model from scratch.[102]

Supervised machine learning needs labeled training data as input for the algorithm to learn to generalize and predict the occurrence of the learned labels on unseen pictures.[103] To illustrate this point, consider a scenario, in which one is tasked with training a model to reliably recognize windows and doors on buildings. In such a case, a substantial number of images depicting buildings in various forms would be required. Subsequently, all doors and windows on the images were to be manually annotated and labeled according to an annotation schema that aligns with the technological specifications of the employed algorithm and the desired model output. The annotation schema for the training dataset constitutes a pivotal design decision. The model's functionality is constrained to the recognition and output of elements that have been explicitly specified. The definition of the information requirements must precede the execution of the study. If the distinction between doors and

---

[98] FLEISCHHACKER David, Object Detection for Improved OCR via Faster R-CNNs on Historical Schematism-State Manuals. Bachelor's Thesis, Graz University of Technology. Graz, pp. 59–62.

[99] KRIZHEVSKY Alex / SUTSKEVER Ilya / HINTON Geoffrey E., ImageNet classification with deep convolutional neural networks. In: Communications of the ACM 60, 6, 2017, pp. 84–90.

[100] KANG Le / KUMAR Jayant / YE Peng et al., Convolutional Neural Networks for Document Image Classification. In: 2014 22nd International Conference on Pattern Recognition: IEEE, 2014, pp. 3168–3172.

[101] DIEM Markus / KLEBER Florian / FIEL Stefan et al., cBAD: ICDAR2017 Competition on Baseline Detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): IEEE, 2017, pp. 1355–1360.

[102] HUSSAIN Mahbub / BIRD Jordan J. / FARIA Diego R., A Study on CNN Transfer Learning for Image Classification. In: Ahmad LOTFI (Ed.) Advances in Computational Intelligence Systems. Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK (= Advances in Intelligent Systems and Computing Ser v.840). Cham: Springer, 2019, pp. 191–202.

[103] SCHÖCH Christof, Information Retrieval. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities, pp. 268–298, here: pp. 289–292.

windows is sufficient for the purpose of the study, then these two categories are adequate. Conversely, if the model is to differentiate between windows of varying shapes, each window shape would constitute a distinct category. This would necessitate a sufficient amount of labeled training data to support each category. The training of a CNN for layout recognition functions analogously. The layout classes that should be recognized must be clearly characterized and continuously labeled in the training data.

Technological progress in the field of artificial intelligence (AI) feeds the hope of largely automatic information extraction from historical documents. In historical research, there are now several examples of the use of AI and computer vision algorithms for the automated processing of source documents. The task commonly called document understanding usually encompasses Document Layout Analysis (DLA) and Handwritten Text Recognition (HTR) for handwritten documents or optical character recognition (OCR) for printed texts respectively.[104]

DLA is a very dynamic research field and there are many promising advancements towards generally applicable solutions. There are tries to task-specifically advance existing computer vision models,[105] there are efforts to understand handwritten documents in a single step, unifying HTR and DLA[106, 107] and promising tries to even include information extraction on top.[108] For printed documents current DLA approaches frequently make use

[104] COQUENET Denis / CHATELAIN Clement / PAQUET Thierry, DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition. In: IEEE transactions on pattern analysis and machine intelligence 45, 7, 2023, pp. 8227–8243, here: pp. 8229–8230.

[105] ZHAO Zhiyuan / KANG Hengrui et al., DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception 2024. Online reference: arxiv.org/pdf/2410.12628v1, (last accessed: 08.02.2025).

[106] HAMDAN Mohammed / RAHICHE Abderrahmane et al., HAND: Hierarchical Attention Network for Multi-Scale Handwritten Document Recognition and Layout Analysis 2024. Online reference: arxiv.org/pdf/2412.18981v1, (last accessed: 08.02.2025).

[107] COQUENET / CHATELAIN et al., DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition, 2023. The DAN model is applied to extract information from 19th century French census tables in the following project: BOILLET Mélodie / TARRIDE Solène et al., The Socface Project: Large-Scale Collection, Processing, and Analysis of a Century of French Censuses 2024. Online reference: arxiv.org/pdf/2404.18706v2, (last accessed: 08.02.2025).

[108] CONSTUM Thomas / TRANOUEZ Pierrick et al., DANIEL: A fast Document Attention Network for Information Extraction and Labelling of handwritten documents 2024. Online reference: arxiv.org/pdf/2407.09103v1, (last accessed: 08.02.2025).

of convolutional neural networks (CNNs)[109] or transformer-based computer vision models.[110, 111]

Conventionally, DLA and OCR/HTR are employed in a sequential manner. Nonetheless, the sequence of steps may be adapted according to the specific circumstances of each use case. Tarride et al. have published a paper on their work on the automatic information extraction from Quebec parish records.[112] Despite the fact that these are handwritten sources, their workflow depicted in Figure 9 reveals that they initiate with a line segmentation task powered by DocUFCN,[113] which is then followed by text recognition. Subsequently, they methodically arrange the disparate text segments into larger, more integrated units and categorize them according to distinct types. This means that the majority of DLA is executed after the completion of the HTR steps. The methodology proves effective for these texts with simple structure. However, problems are expected to arise when attempting to implement it in the context of highly fragmented pages, as exemplified by the pages from the Schematismus.

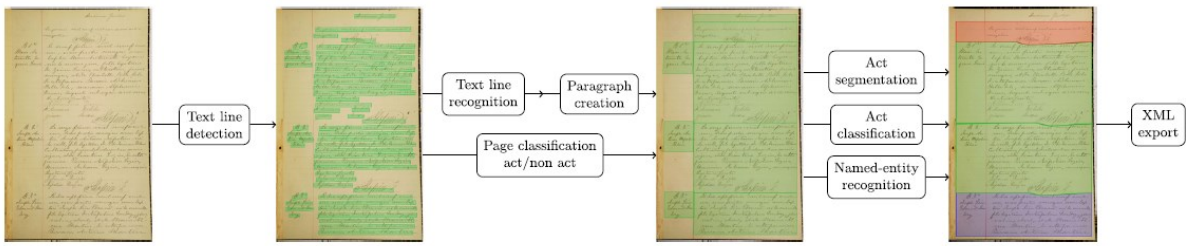


Figure 9: Graphic from Tarride et al. illustrating their sequence of procedural steps.[114]

The extraction pipeline from the Schematismus deliberately performs layout segmentation before line detection and OCR. In a recent publication the team of the Schematismus Project explains how they used large amounts of synthetic Schematismus pages to train a

[109] SHEIKH Talha Uddin / SHEHZADI Tahira et al., UnSupDLA: Towards Unsupervised Document Layout Analysis 2024. Online reference: arxiv.org/pdf/2406.06236v1, (last accessed: 08.02.2025).

[110] SHEHZADI Tahira / STRICKER Didier et al., A Hybrid Approach for Document Layout Analysis in Document images 2024. Online reference: arxiv.org/pdf/2404.17888v2, (last accessed: 08.02.2025).

[111] WANG Jiawei / HU Kai et al., DLAFormer: An End-to-End Transformer For Document Layout Analysis 2024. Online reference: arxiv.org/pdf/2405.11757v1, (last accessed: 08.02.2025).

[112] TARRIDE Solène / MAARAND Martin / BOILLET Mélodie et al., Large-scale genealogical information extraction from handwritten Quebec parish records. In: International Journal on Document Analysis and Recognition (IJDAR) 26, 3, 2023, pp. 255–272.

[113] BOILLET Melodie / KERMORVANT Christopher et al., Multiple Document Datasets Pre-training Improves Text Line Detection With Deep Neural Networks 2020. Online reference: arxiv.org/pdf/2012.14163, (last accessed: 08.02.2025).

[114] TARRIDE / MAARAND et al., Large-scale genealogical information extraction from handwritten Quebec parish records. p. 260.

Faster R-CNN model.[115] The synthetic training pages were produced, by making use of a recreated Schematismus font. This approach aided both the layout and text recognition models. Performing layout detection first and feeding the individual layout snippets to a Tesseract OCR model,[116] which was itself fine-tuned with the artificially created font, was able to lower the character error rate (CER) by 71.98 % when compared to an out-of-the-box Tesseract OCR model.

The artificially generated training pages were used for a first training iteration of the Layout Model. This first model was used to label roughly 1000 authentic Schematismus pages. These pages were then controlled and corrected manually to generate a gold standard dataset for any further model training.

Since the publication of the aforementioned paper by Fleischhacker et al., the project team further optimized the technological basis of the layout model by first moving on to a YOLOv9[117] model and then to a Co-DETR transformer-based architecture.[118] The principal workflow of performing layout detection, sorting the layout elements, performing OCR on every layout element and exporting the text and layout information remains the same.

The output of this pipeline is a dataset that contains information on the location of the bounding box of each layout element, its predicted layout category and its contents in the form of the transcribed text. Figure 10 shows a visualization of the predicted bounding boxes of layout elements of a short section from the Schematismus. Different colors indicate different layout categories, and each bounding box encompasses a single entity of the respective layout category.

---

[115] FLEISCHHACKER / GOEDERLE et al., Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach, pp. 9–15.

[116] SMITH R., An Overview of the Tesseract OCR Engine. In: Proceedings / Ninth International Conference on Document Analysis and Recognition. Curitiba, Paraná, Brazil, September 23 - 26, 2007. Los Alamitos, Calif.: IEEE Computer Soc, 2007, pp. 629–633.

[117] WANG Chien-Yao / YEH I-Hau et al., YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information 2024. Online reference: arxiv.org/pdf/2402.13616v2, (last accessed: 08.02.2025).

[118] ZONG Zhuofan / SONG Guanglu et al., DETRs with Collaborative Hybrid Assignments Training 2022. Online reference: arxiv.org/pdf/2211.12860v6, (last accessed: 08.02.2025).

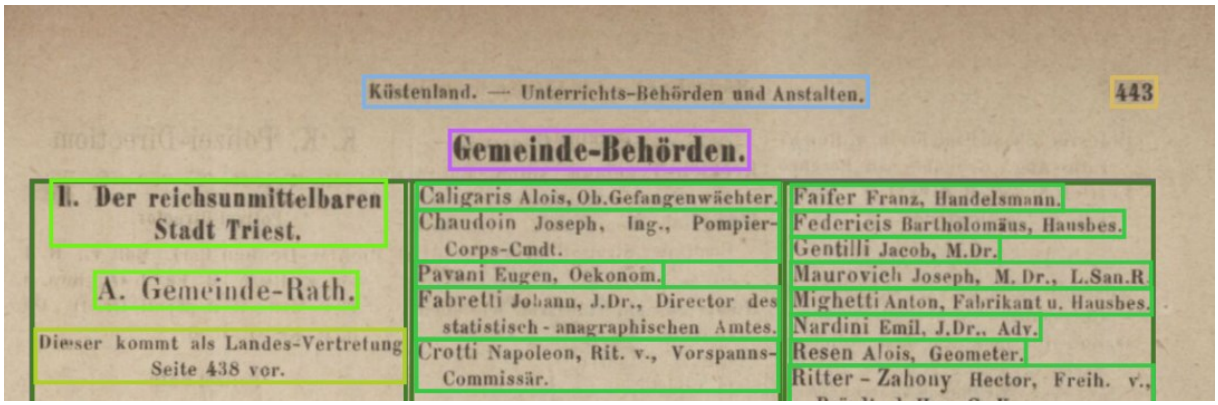


Figure 10: Predictions of the layout model (Co-DETR) visualized using the Bounding Box Editor.

The remaining chapters of this thesis will explain, analyze and evaluate the output of this specific Layout Model. In order to do so, it is necessary to analyze the data representation in the printed Schematismus and how the project team of the Schematismus Project developed an annotation schema that enables the Layout Model to automatically extract the basic components of said data representation.

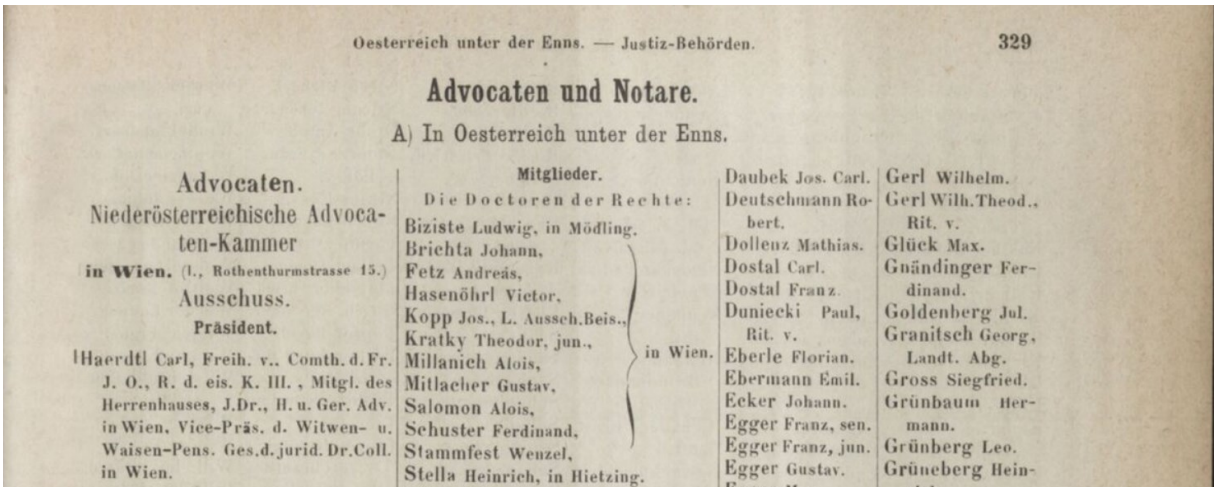
The following discussion will proceed in three steps. First, the data representation of the Schematismus will be analyzed. This will be followed by a description of the abstraction of the layout components into an annotation schema. Based on these discussions, a basic data model for the translation of the Schematismus into a digital database will be deducted.

4.2 The Data Representation of the Schematismus

As stated in Chapter 2.2, the Schematismus's layout underwent a period of consolidation in its final phase. In this regard, no significant changes are observed from the late 1860s until the dissolution of the monarchy. In addition, from 1876 through 1918, the Schematismus was published on an annual basis. All these editions have been preserved and made digitally available at the ÖNB, thus constituting the longest continuous series of consecutive editions of the entire series. When the increased number of individuals per edition is also taken into account, these last years hold a disproportionately large amount of the entire accessible dataset.

Consequently, the Schematismus Project team has identified these years as priority targets. The initial project phase will entail the extraction and creation of a database based on these later editions. All considerations in the following sections are with respect to the peculiarities of their specific layout.

Conceptually, the Schematismus's layout makes a fundamental distinction between individuals and institutions as the basal units of information.[119] Individuals are always assigned to institutions. Institutions are formally hierarchical, with the hierarchical dependencies encoded via layout. A screenshot from Page 329 of the 1876 Schematismus can be utilized to systematically decompose the structural tree that is efficiently represented by these typographic configurations.



Oesterreich unter der Enns. — Justiz-Behörden. 329

Advocaten und Notare.

A) In Oesterreich unter der Enns.

Advocaten.
Niederösterreichische Advocaten-Kammer
in Wien. (I., Rothenthurmstrasse 15.)
Ausschuss.
Präsident.
Haerdtl Carl, Freih. v., Comth. d. Fr. J. O., R. d. eis. K. III., Mitgl. des Herrenhauses, J.Dr., H. u. Ger. Adv. in Wien, Vice-Präs. d. Witwen- u. Waisen-Pens. Ges. d. jurid. Dr. Coll. in Wien.

Mitglieder.
Die Doctoren der Rechte:
Biziste Ludwig, in Mödling.
Brichta Johann,
Fetz Andreas,
Hasenöhrl Victor,
Kopp Jos., L. Aussch. Beis.,
Kratky Theodor, jun.,
Millanich Alois,
Mitlacher Gustav,
Salomon Alois,
Schuster Ferdinand,
Stammfest Wenzel,
Stella Heinrich, in Hietzing.
} in Wien.
Daubek Jos. Carl.
Deutschmann Robert.
Dollenz Mathias.
Dostal Carl.
Dostal Franz.
Duniecki Paul, Rit. v.
Eberle Florian.
Ebermann Emil.
Ecker Johann.
Egger Franz, sen.
Egger Franz, jun.
Egger Gustav.
Gerl Wilhelm.
Gerl Wilh. Theod., Rit. v.
Glück Max.
Gnändinger Ferdinand.
Goldenberg Jul.
Granitsch Georg, Landt. Abg.
Gross Siegfried.
Grünbaum Hermann.
Grünberg Leo.
Grüneberg Hein-

Figure 11: Top section of page 329 of the Schematismus from 1876.[120]

Figure 11 shows a section from the regional administration of the "Erherzogtum Österreich unter der Enns", now Lower Austria. As can be seen from the page header, this is an excerpt from the "Justiz-Behörden" (judicial authorities). The section's main heading reads "Advocaten und Notare" (lawyers and notaries). The following, slightly smaller subheading makes a restriction that only individuals "In Oesterreich unter der Enns" are listed. A peculiarity that complicates the accurate modeling of the offices is the fact that the section of the "Justiz-Behörden" in Lower Austria also encompasses the judicial authorities of Upper Austria and Salzburg. Judging by the listings of other authorities and comparing them with other crown lands, this seems to be a peculiarity of the judicial sector. Nevertheless, this circumstance necessitates the aforementioned specification through the subheading. The "Advocaten" were organized in the "Niederösterreichische Advocaten-Kammer", which

[119] This combination of the personal and the institutional level is also foundational in modern visualizations of organizational structures, such as organizational charts. Compare: SCHEWE Gerhard, Organigramm 2018 (=Gabler Wirtschaftslexikon). Online reference: wirtschaftslexikon.gabler.de/definition/organigramm-44460/version-267771, (last accessed: 01.02.2025).

[120] Digitized by the ÖNB and publicly accessible under: https://alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1876&page=490, (last accessed: 06.01.2025).

had its headquarters at Rotenturmstrasse 15 in Vienna and was headed by an “Ausschuss” chaired by Freiherr Carl Haerdtl as “Präsident”.

These few lines allow the extraction of several information units that can be linked to Carl Heardtl. However, Carl Haerdtl’s embedding in the apparatus of the civil service is not conclusively represented, since the judicial authorities of Lower Austria do not represent the highest office of the state. The higher ranks of the monarchy’s administration must be collected from other locations within the Schematismus and correctly referenced. The Schematismus offers its table of contents and the headings of the preceding 328 pages as references to conclusively traverse the hierarchical ladder to the top of the state. It can be deduced from the table of contents that the administration of Lower Austria was one of several “Vertretungen und Verwaltungen der einzelnen Königreiche und Länder”.[121] To ensure the proper orientation of the Cisleithanian regions within the organizational chart of the state administration, it is imperative to consult all major headings provided in the primary section of the Schematismus.

Therefore, the integration of a single listed person into a rudimentary schematic network of the Habsburg state administration, as illustrated in Figure 12, necessitates careful examination and substantiated assumptions. The automatic compilation of this institutional network for each entry in the print version of the later editions of the Schematismus would constitute a significant achievement for the Schematismus Project. However, several complicating factors must be further considered. It is not uncommon for an individual to possess multiple entries in the Schematismus. All these entries should be correctly mapped to the same individual. As has been previously indicated, the Schematismus, for that purpose, provides an extensive register of names.

[121] The official name of the Cisleithanian provinces is commonly translated as "The Kingdoms and Lands represented in the Imperial Council".

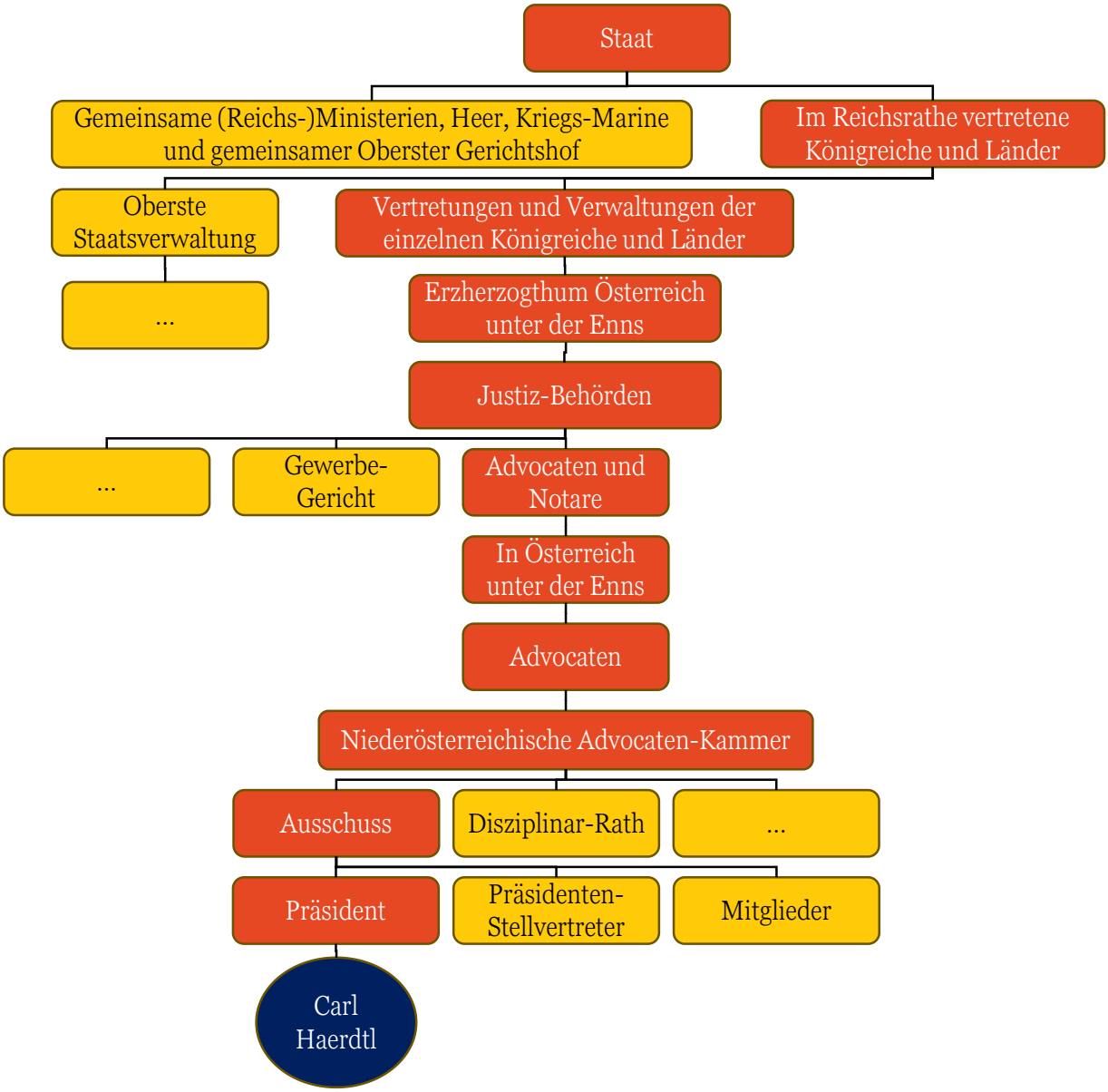


Figure 12: Schematic depiction of a hierarchical office structure that can be inferred from a single entry for a Baron Carl Haertl in the 1876 Schematismus (illustration by the author).

A preliminary investigation into "Carl Haerdtl" in this list of names initially yields no result, as the person is not listed with this spelling. However, there appears to be a number of individuals bearing the surname "Härdtl". The search yields a "Härdtl Carl" and a "Härdtl Carl, Freih. v.". The existence of separate listings indicates that these are distinct individuals. It is noteworthy that neither of the aforementioned sources makes any reference to page 329. According to the source, Freiherr von Härdtl is mentioned on pages 91, 119, 328, 852, and 859. A swift comparison with page 328 substantiates the assumption that this is likely an error in the name register. Freiherr Carl "Haerdtl" is most likely identical to Freiherr Carl "Härdtl" and reference should be made to page 329 instead of 328.

In the final database, the entry “Carl Haerdtl” presented in Figure 12 would have to be linked with all the other entries that exist for Freiherr Carl Härdtl according to the name index. The same individual was also the recipient of several medals, including the “Orden der eisernen Krone” and the “Franz-Joseph-Orden”. In addition, he served as a member of the administrative council of the “K. K. priv. österr. Staatseisenbahn-Gesellschaft” and vice president of the “Witwen- und Waisen-Pensions-Gesellschaft des juridischen Doctoren-Collegiums”.

The complexity of automating the task of reconstructing hierarchical and cross-page relationships is aptly illustrated by the fact that several errors in the print were discovered when attempting to trace a randomly selected individual within a single year of the Schematismus. As demonstrated by the provided example, error-free text recognition and very detailed layout segmentation are indispensable for the automated reconstruction of organizational affiliations. However, these competencies alone are insufficient to ensure the accuracy of the results. Solutions that compare the page extracts with the table of contents and name index must be robust enough to understand different spellings and be able to correct errors such as incorrectly referenced pages. Minor scanning errors, such as the superfluous vertical line preceding “Haerdtl” in Figure 11, should also be processed without issue. Many of these issues concern later project phases. The focus of the next chapter is the bespoke annotation schema that was developed by the project team of the Schematismus Project, and which should enable the Layout Model to automatically recognize and extract the basic entities that construct organizational networks like the one of Carl Haerdtl.

4.3 The Annotation Schema for the Layout Segmentation Model

While machine learning models offer powerful recognition capabilities, they remain fundamentally limited to identifying only those visual elements, explicitly specified in their training data. Consequently, training data design constitutes a critical component of data modeling for the Schematismus Project, as it predetermines the nature and quality of information available for database input. The annotation schema acts as an intermediary between the complexity of the data representation of the print described in the previous chapter and the recognition capabilities of computer vision technologies introduced in chapter 4.1. It must strike a balance between generalization and granularity of categories to allow both uniform recognition and detailed organizational reconstruction.

By way of establishing different categories, the training data defines the level of detail of the returned layout information. For instance, two categories would be enough to train the computer vision model to separate all headings from the paragraphs. But many more, well characterized categories are needed to distinguish different headings and recognize columns, header and footer areas. Given the extensive scope of the Schematismus, it is imperative to comprehend the recurrent typographical patterns and idiosyncrasies that characterize it.

With regard to the complex layout of the primary source, the team of the Schematismus Project therefore decided to focus on a standardized annotation of visual elements. Elements that look similar are classified in the same way, regardless of the meaning they actually have in the source context. The semantic assignment is only carried out in a post-processing step. This ultimately leads to a high-performance layout detection model that can recognize the central classes in the layout - the paragraphs - with very high accuracy.

The subsequent section delineates the components of the annotation schema that was used to produce the training data for the Layout Model. The project team utilized the freeware Bounding Box Editor to annotate the layout elements by drawing rectangular bounding boxes on the scanned pages.[122]

4.3.1 Individuals

As opposed to institutions, individuals necessitate a reduced number of classes to accommodate their diverse stylistic appearances. Individuals are represented by left-aligned text positioned beneath the respective headings. According to the annotation schema there are two types of layout elements that encode individuals, namely the 'Paragraph' and the 'BigParagraph':

- Paragraph: The 'Paragraph' constitutes the most basic structural element that is recorded by the annotation schema. Typically, a 'Paragraph' is considered to represent one individual. The content of the 'Paragraph' encompasses the name of the individual, along with, in numerous instances, supplementary information, including noble and official titles. From a visual perspective, the 'Paragraph' is characterized by its left-aligned text, which is integrated into the standard

[122] https://github.com/mfl28/BoundingBoxEditor (last accessed: 02.02.2025)

column structure that is characteristic of the Schematismus. A perusal of the Schematismus reveals that the font size of 'Paragraphs' is consistent across all pages of the examined edition. The first word of each 'Paragraph' is printed with a slightly larger font size and, when space permits, with increased inter-letter spacing. 'Paragraphs' can be single or multi-line, with all but the first line being slightly indented. In order to achieve the desired information gain, it is imperative that all 'Paragraphs' be distinctly separated from one another. This method is the only viable approach for recording all individuals within the administrative structure in a systematic and comprehensive manner. In the case of multi-line 'Paragraphs', the indenting of subsequent lines facilitates the identification of a clear visual break. In the case of several consecutive single-line 'Paragraphs', the distinguishing criteria are the minimal difference in size of the first word and the lack of indentation of the following line. As illustrated in Figure 13, the Layout Model effectively resolves both cases in this particular excerpt. The 'Paragraphs' are demarcated by dark yellow boxes. The initial 'Paragraph' is comprised of two lines, while the subsequent 'Paragraphs' each consist of one line.

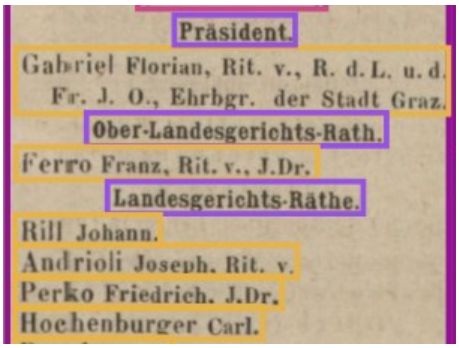


Figure 13: Paragraphs are marked by dark yellow bounding boxes.

- BigParagraph: In regard to the composition of their contents, 'BigParagraphs' are analogous to 'Paragraphs', as they each comprise a single individual. In terms of style, a notable distinction emerges: 'BigParagraphs' extend to the full width of the page, thereby occupying the entire available space. Consequently, the presence of individuals in 'BigParagraphs' is indicative of their significant roles within the organizational hierarchy. At the same time, 'BigParagraphs' and 'Paragraphs' exhibit numerous similarities. Both elements are aligned to the left and employ the same font, though the font size in 'BigParagraph' is marginally larger. The differences are clearly visible in Figure 14. The 'BigParagraph' is

Figure 14: The 'BigParagraph' category is marked by the blue bounding box.

outlined in blue and contains the personal attributes of the President of the "K. K. Ober-Landesgericht" for Styria, Carinthia and Carniola. The subordinate offices are then arranged in columns, starting with the vice president.

4.3.2 Institutions

The depth of the hierarchy of the monarchy's administrational structure reveals itself and can be reconstructed via the manifold of headings. While only two different layout categories are needed to encompass all possible variations of how individuals can appear, the annotation schema would require a multitude of categories to achieve the same for the various headings. By assigning them to five different categories, the project team paid tribute to the goal of retrieving additional hierarchical information, while at the same time keeping the categories distinguishable according to shared visual characteristics. The names of the five classes (Ho, 'H1', 'H2', 'H5' and 'H6') indicate that the definition of the annotation schema was an iterative process. Some classes ('H3', 'H4') did not have class-wide features that could be uniformly replicated by the Layout Model and therefore had to be incorporated into other categories.

Figure 15: Excerpt from a page that was annotated by The Layout Model according to the Schematismus Project's annotation schema. The excerpt shows all common heading categories. Ho-turquoise, 'H1'-red, 'H2'-magenta, 'H5'-green, 'H6'-pink.

The most striking common distinguishing feature of all headings is that they are always centered horizontally. The screenshot in Figure 15 shows the highest offices of the

“Herzogthum Steiermark” (Styria) and the cascade from the “Landes-Vertretung” to the “Landes-Ausschuss” and further down the hierarchy. The extract was chosen because it contains all five heading categories in a small space.

- ‘H0’ is the largest heading known to the annotation schema. It only refers to page-wide headings. In Figure 15, the layout model has assigned “Herzogthum Steiermark.” to category ‘H0’. In the editor, the heading is highlighted in turquoise.

- H1: The category ‘H1’ is similarly exclusively assigned to large, page-wide headings. ‘H1’ headings are visibly smaller than ‘H0’ headings, and ‘H0’ headings tend to appear only at the top of pages. What is important is the reliable separation and correct order of the headings, which is usually provided by the Layout Model, as shown in Figure 15.

- H2: Headings in the ‘H2’ category are very common. The Layout Model designates the smallest page-wide headings, as well as larger headings in the columns (see Figure 15, green boxes) as ‘H2’ headings. The ambiguity regarding the placement of ‘H2’ is likely attributable to corresponding ambiguous annotations. It appears that, at this stage, the decision to merge ‘H2’ and ‘H3’ into a single category, which was made in order to address the issue of unreliable results, could be reversed ex post by assigning all ‘H2’ headings within columns to the ‘H3’ category.

- H5: Headings in the ‘H5’ category are the lowest-ranking headings in the annotation schema. ‘H5’ headings only appear in columns. They are usually the first heading above a ‘Paragraph’. In Figure 15 they are marked in purple.

- H6: The ‘H6’ category includes all text that is centered and enclosed in parentheses. Most of the time, these lines don’t semantically serve the purpose of a heading. They usually add additional information to the previous heading, such as addresses or, as in the right column of Figure 15, municipalities that belong to respective districts.

The very fact that the annotation schema is much simpler than the complexity and depth of the source means that major compromises are made, especially in recognizing the headings. The team of the Schematismus Project made the decision to prioritize the uniform recognition of similar structural elements and thus stable layout recognition. The aim

was to create a model that recognizes a small number of classes with a high degree of accuracy over several yearly editions of the Schematismus. In the interest of automating the extraction process, a design decision was made to largely dispense with the inclusion of historical meaning content in this process step and to shift this aspect to the downstream post-processing step as far as possible.

The depictions in Figure 15 are suitable to illustrate the high degree of reliability of the Layout Model, although they also reveal some recognition errors. The 'H5' label in the first line of the middle column should be a 'Paragraph', and the singular 'H5' label in the right column should be 'H6'. It was mentioned that the Layout Model is continuously improved by the project team. The predictions visible in Figure 15 were produced by a YOLOv9 architecture, which was the best performing model in July 2024. Letting the same section be annotated by the newer CoDetr model yields more consistent results and shows fewer errors.

4.3.3 Additional Layout Information

While paragraphs and headings allow the deconstruction of major parts of the Schematismus's contents, further layout components need to be annotated to facilitate downstream tasks to reconstruct the reading order and capture peculiarities of the layout.

The editors of the Schematismus tried to save space – supposedly to save printing costs[123] – which led to numerous peculiarities of the format, such as abbreviations or special characters. These additionally complicate automated information extraction. A typical example are curly brackets. They are utilized to assign properties to a number of individuals. A good example of how curly brackets are used in the layout can be found in the screenshot shown in Figure 11 on page 40. The middle column encompasses members of the "Niederösterreichische Advocaten-Kammer". Mödling, the place of office of the first of the "Doctores", Ludwig Biziste, is appended to the name information, separated by a comma. The next ten individuals listed are based in Vienna. In this case, the location is assigned collectively using curly brackets.

Functionally the bracket assigns the property to its right to all elements on the left side. The Layout Model resolves this by making the right element a heading for the left

[123] Schlögl claims, for example, that the Austrian Biographical Dictionary, another expansive printed register, was optimized for printing costs: SCHLÖGL / LEJTOVICZ, Die APIS-(Web-)Applikation, das Datenmodell und System, 2020, p. 31.

paragraphs. For many instances of curly brackets, the difficulty lies, for one, in determining the exact extent of the bracket and then in identifying the respective property for those entries outside the bracket. Both problems can be illustrated using the same example illustraded in Figure 11. A trained computer vision model should be able to reliably detect the curly bracket in the middle column. Using only visual information it might be difficult to tell, whether Wenzel Stammfest is still covered by the bracket. However, identifying the place of work for all individuals outside the bracket is a different task entirely. Thefore, both layout and semantic text information are needed to correctly interpret sections with curly brackets.

The frequent occurrence of curly brackets in the Schematismus caused the project team to introduce a separate 'Curly' layout class.

- Curly: As exemplified by Figure 16, the 'Curly' class is used to frame all textual components that are affected by the curly bracket. The 'Curly' container (red) encompasses the 'Paragraphs' to the left of the bracket and the 'H5' heading to its right.

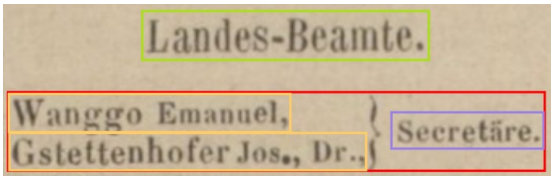


Figure 16: Prediction of a curly container (red). The element to the right of the curly bracket is labeled as H5.

Other layout components are categorized and extracted to facilitate downstream tasks. The division of a page in tables and columns is needed to algorithmically reconstruct the reading order of layout elements on a page. Page number and page header, on the other hand, can be used to establish connections across different sections of the Schematismus utilizing both the page references in the table of contents and the name register.

- Table & Column: The purpose of the 'Table' category is to identify those areas of the layout that are organized in columns. The objective is not to identify specific table formats, as these are not present in the Schematismus. Instead, the focus is on establishing a preliminary classification into page-wide sections and those arranged in columns. The regions are delineated by a box that fully encompasses all the columns. The columns – typically three, although occasionally two – are individually provided with a bounding box, which is labeled 'Column'.

- Nested-Table & Nested-Column: 'Nested-Tables' and 'Nested -Columns' are used when an additional small two-column table is integrated into a standard column of the layout. The annotation is analogous to their big siblings. The 'Nested-Tables' cover the entire area, whereas 'Nested-Columns' only extend to cover the individual columns. An example of the use case is shown in Figure 17. The teachers of the "K. K. Ober-Gymnasium" in Maribor are catalogued in a 'Nested-Column' – presumably to save space – and are demarcated in blue by the Layout Model. The illustration also shows that the 'Nested-Table' is highlighted in pink and the entirety of the structure is contained within the 'Column' (black) and 'Table' (dark blue) elements.

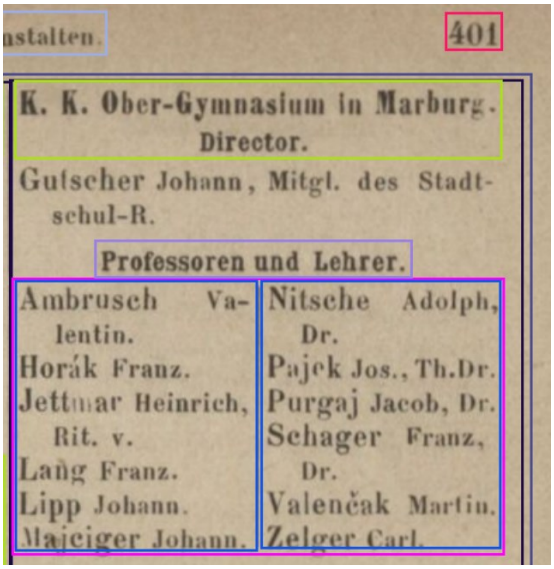


Figure 17: Nested-Columns (blue) and Nested-Table (pink).

- Page Number & Header: In newer editions of the Schematismus, the page number is consistently located in the upper corner of pages, the 'Header' consistently in the center at the top of most pages. Both categories are readily identifiable by virtue of their layout as isolated elements and their contents will be of use in terms of future referencing, specifically with regard to linking to the table of contents and name register.

4.4 A Model for the Extracted Data

Having examined the structure and granularity of the layout categories that were defined by the project team in their annotation schema, the present master's thesis attempts to translate these predefined categories into a conceptual framework. The objective of this endeavor is to ensure that the conceptual framework accurately represents the organizational hierarchies documented in the Schematismus. This transition from extracted layout elements to a coherent data model represents a critical methodological step in the present digitization approach. The challenge lies in determining, whether the layout-based classification system can, independently of semantic text analysis, provide sufficient structural

information to reconstruct the administrative hierarchies that the Schematismus was designed to document.

There is still more information contained in the Schematismus, which could be unlocked by semantically analyzing the bounding box contents on the textual level. This further step is very relevant to decode the entire Schematismus data and will be performed in later project phases. However, layout information alone should give a sustainable and coherent backbone for the database, connecting individuals and institutions in an organizational framework.

Drawing upon the established approaches to prosopographical data representation discussed in chapter 3, three principal entity types are identified as constituting the fundamental components of the Schematismus's organizational structure: 'Person', 'Position', and 'Institution'. In contrast to more intricate prosopographical systems such as APIS, which encompass exhaustive biographical data, this current model concentrates exclusively on administrative relationships that can be deduced from layout components.

This tripartite structure partially aligns with the categories of the annotation schema presented in the previous chapter. The 'Person' entity will be filled by the contents of 'Paragraph' and 'BigParagraph' layout categories, while the 'Institution' entity will draw on the contents of the various heading categories. The 'Position' entity was introduced to capture the official roles or titles that mediate between individual officers and institutional structures.

The structuring elements of the print layout, especially the different types of headings, depict a classical organizational structure, including institutions, positions and persons, as typically comprised in an organizational chart. The data model should reflect this organizational structure and be able to visualize the organizational chart as closely to the print template as possible.

In order to illustrate the relational structure of the data model under consideration, it is proposed that the relationships between the three entity types be formalized as semantic triples. To illustrate this formalization process, the example from the previous chapter 4.2 (Figure 11) is readopted. The following types of relational statements pertaining to Carl Härdtl can be extracted:

1. Person-Position Relationship (A named individual holds a specific administrative position):

$$\underset{Position}{< \text{Präsident} >} \quad \underset{Relation}{< \text{is occupied by} >} \quad \underset{Person}{< \text{Carl Härdtl} >}$$

2. Position-Institution Relationship (An administrative position exists within a specific institutional context):

$$\underset{Institution}{< \text{Ausschuss} >} \quad \underset{Relation}{< \text{has} >} \quad \underset{Position}{< \text{Präsident} >}$$

3. The Institution-Institution Relationship (An administrative unit exists within a larger institutional hierarchy):

$$\underset{Institution}{< \text{Niederösterreichische Advocaten} - \text{Kammer} >} \quad \underset{Relation}{< \text{has} >} \quad \underset{Institution}{< \text{Ausschuss} >}$$

These three relationship types capture the fundamental organizational structure documented in the Schematismus, with the 'Position' entity functioning as a pivotal mediating element between individuals and the institutional framework. The 'Position' entity functions in two ways. It both characterizes individuals within the bureaucracy and defines functional roles within organizational units. A 'Position' like "Präsident" can at the same time be attributed to an individual but also occupy a functional role in the organizational tree. Much more for example than "Ausschuss", which is clearly impersonal and belongs to the institutional structure. The logic of the above triples dictates that 'Positions' can be occupied by 'Persons', 'Institutions' can have one or more 'Positions' and 'Institutions' themselves can have one or more other 'Institutions'. This last connection between 'Institutions' indicates hierarchical relationships, just as they are apparent in the printed Schematismus.

The Layout Model for the Schematismus Project yields a reliable separation of 'Persons' and 'Institutions'. Nevertheless, several methodological challenges must be addressed before the semantic triple structure outlined above can be fully populated for the entire Schematismus. In particular, it is necessary to ascertain whether the layout-based classification system currently employed can reliably establish directional relationships between entities at different hierarchical levels. If this is not possible, we need to understand what kind of information is needed additionally and how to best obtain it. The close connection between the print and the digital implies that this evaluation must be based on a

thorough understanding of the data logic of the printed source. The understanding of this data logic allows us to evaluate the sufficiency of the extracted data and determines the quality of further process steps.

Furthermore, it must be evaluated how best to fill the ‘Position’ entity. The titles and roles in the administrative structure, which constitute the ‘Positions’ can only be partially reconstructed by use of layout recognition. The pertaining information is not always printed as headings. In the case of less important official titles than a state governor, it is often appended in the ‘Paragraph’ following after the individual’s name. As this information is not accessible without analyzing the text semantically, this evaluation must largely be postponed to later project stages. Consequently, the evaluations in the subsequent chapter will primarily focus on the establishment of directional relationships between the extracted entities.

5 Evaluating the Sufficiency of the Extracted Data

The three semantic triples that constitute the preliminary data model for this study build on a standard subject-predicate-object structure. Although the complications relating to the 'Position' entity have been discussed, the subject and object entities can be largely populated utilizing the extracted layout elements. However, it is more difficult to deduct the predicate that encodes the nature of the relationship between the respective entities from the extracted information.

The direction of the relationship is needed to establish hierarchical superiority or inferiority of entities within the organizational structure that is represented by the Schematismus. Because the annotation schema is only an abstraction of the complexity of the printed source, the structured data output of the Layout Model doesn't represent the full depth of information needed for a straightforward reconstruction of directed relationships. The sequence of elements partially informs which entities should be linked, whereas the predicted layout class can expand the information in this regard and hint at the direction of the connection, as far as the five-part classification scheme allows.

This chapter 5 explores methodological challenges and technical solutions for reconstructing the organizational hierarchy embedded in the Schematismus layout. The objective is to ascertain whether the layout data that has been extracted contains sufficient information to reconstruct administrative hierarchies without necessitating an extensive semantic analysis of the text content. To support this effort the project team supplied a dataset containing the structured data output of the Layout Model for an entire edition of the Schematismus.

5.1 The CSV Dataset of the Schematismus from 1878

The Schematismus edition of 1878 belongs to the continuous series of successive editions that has been identified as the primary target of the initial project phase of the Schematismus Project. With a view to evaluating the capabilities of their extraction pipeline the project team employed the Layout Model to predict layout categories for all pages of this specific edition. At the time this experiment was performed in July 2024 the best performing version of the Layout Model was based on a YOLOv9 architecture.

This specific model had been trained on circa 1000 labeled Schematismus pages from various editions. Although the Layout Model is continuously evolved all further evaluations

in this thesis will be made with regard to the structured output of the aforementioned version. Nevertheless, this approach is expected to produce valuable insights, since the layout will always be represented by the same annotation classes, as long as the annotation schema doesn't change. The accuracy of predictions, however, is expected to increase with every evolutionary upgrade.

The structured output of the Layout Model was received as a data dump in the form of 1126 CSV files, one file per scanned page. CSV stands for "comma separated values". The CSV format is a widely used standard and encodes data in a classic table structure. Each file contains the layout predictions for a single page in tabular form, with each row corresponding to a separate layout element identified in the source material. For each row the pertaining information is systematically organized across several dimensions. As illustrated by the color-coded extract in Figure 18, the files adhere to a standardized format, with each data component clearly separated into distinct columns.

A column labeled "class" contains the information regarding the predicted layout category for the respective layout element, while the location of the same on the scanned source page is encoded via coordinates in four of the other columns. The last column contains the textual contents of the respective layout element as transcribed by the custom OCR model of the Schematismus Project.

Figure 18: A snippet from the CSV-data that serves as input for the reconstruction of the organizational hierarchy. The columns are color-coded. The first row shows the column labels.

It must be noted that the location and class label for each layout element were predicted by the Layout Model, the sequence of layout elements within the CSV table, however, was determined in a post-processing step by a custom resorting algorithm. The algorithm was

developed by the project team to represent the output of the Layout Model according to the reading order of the source document. The sequence of elements takes into account columns, nested columns or sections with curly brackets.

In the larger context of the Schematismus Project, the data sample from 1878 represents a fraction of the data that will undergo further post-processing as soon as the project proceeds to the stage of mass data extraction. The OCR-transcribed text content will then be processed by named entity recognition (NER) in order to extract semantic information with a view to further enriching the dataset. The focus of the present analysis remains on reconstructing hierarchical relationships between the extracted layout elements. The next chapter is dedicated to reconstructing as much of the hierarchy as possible for a limited set of pages by use of the layout classes and limited, easy to retrieve context information.

5.2 Hierarchy Reconstruction

The example in chapter 4.2 has demonstrated how the various headings of the printed data representation can be assembled in a hierarchical network structure. In said example, Carl Haerdtl, the individual contained in a 'Paragraph' layout element, was depicted in an organizational chart, displaying all direct institutional dependencies from his position as "Präsident" of the "Ausschuss" of the "Niederösterreichische Advocaten Kammer" to the top of the administration of the Habsburg monarchy.

The chart and all containing connections were produced manually. However, in accordance with the targets of the Schematismus Project, it should be evaluated whether this process can be automated based on the extracted information. As a first step it is helpful to formalize the rules by which hierarchical relationships are represented in the printed data.

Within the organizational framework of the printed Schematismus, the hierarchical relationships between administrative units are encoded through two complementary mechanisms: typographical differentiation and sequential positioning. How typographical differentiation creates visual indications of relative importance between headings has been discussed extensively. The role of sequential positioning, on the other hand, has so far not been discussed with the same level of detail. However, the sequential placement of headings within the document's reading order is similarly important.

For the printed data representation, the general rule applies that, when a larger heading is followed by a smaller one, a direct hierarchical relationship is observed to be

established. In such cases, the smaller heading becomes subordinate to the larger one. Conversely, when a smaller heading is followed by a larger one, this indicates a hierarchical break rather than a direct relationship. In such cases, both headings are respectively subordinate to the nearest larger heading in reverse reading order. This phenomenon can be conceptualized as the establishment of "zones of influence" within the document structure, whereas each heading exercises administrative authority over all subordinate elements until another heading of equal or greater prominence appears.

These zones of influence can be nested within one another, creating a complex hierarchical tree. A 'H1' heading can span over several 'H2' headings, which themselves might have their own influential zones spanning 'H3' headings. Conceptually, this nested structure is analogous to a hierarchical markup language, wherein the opening tags signify the commencement of a hierarchical zone, while the closing tags denote its termination (<H1_area> <H2_area> <H3_area> … </H3_area> <H3_area> … </H3_area> </H2_area> <H2_area> <H3_area> … </H3_area> </H2_area> </H1_area>, where "/" signals the closing tag).

The question arises, to what extent these rules can be applied to the extracted digital data and whether the zones of influence of headings can be reconstructed. The CSV-tables list all heading elements in the correct reading order and assign them to a limited number of hierarchical categories. The position of each heading's listing indicates only the opening tag of its zone of influence. Analogous to the rules formalized for the printed data representation, it can be deduced that the zone of influence of any heading ends whenever another heading that is of the same or of a higher hierarchy level is reached. The zone of influence of a 'H3' heading ends, when there is another 'H3' heading, or a 'H2' heading, whichever comes first within the order of the CSV-file.

To slightly reduce the complexity of the task an initial set of rules was deduced and implemented, that focuses exclusively on the sequential positioning of layout elements within the CSV-tables. Paragraphs and headings demand different sets of rules, as their sequential positioning has different respective implications. Each paragraph following a heading is subordinate to that heading, while directly subsequent paragraphs share the same hierarchical level. For sequential headings a different logic must be applied. If a heading is immediately followed by another, this indicates a disruption in structure and hierarchy. Sequential primacy indicates hierarchical superiority. Headings, on the other hand,

that contain one or more paragraphs, and are thus not in a direct sequential order, are expected to belong to the same hierarchical level.

Building on this rudimentary set of rules an attempt was made to reconstruct as much of the hierarchical information as possible. For this experiment, the Python library Pandas[124] was utilized to manipulate the tabular CSV-data. A commented version of the code is annexed to this thesis after the Bibliography as "Annex 1: Commented Code Repository". The code presented there has undergone an iterative development process. The objective was to initiate the process with a single Schematismus page with minimal layout complexity. After the efforts proved successful on that page, it was decided to proceed and attempt to implement the model on different pages. The code was subjected to three such iterative major revisions to reach its current form, outlined in the annex. The scope of this experimental implementation was narrowed down to two sections of the Schematismus of 1878, namely the pages displaying the court of Emperor Franz Joseph II and the section relating to the regional administration of Styria. The attempt was initiated on page 18 of the Schematismus of 1878, which displays a segment of the court of Emperor Franz Joseph II (see Figure 19)[125]. The depiction of page 18 will serve to illustrate the logic by which the code was designed.

The presented reconstruction approach disregards the predicted heading categories. As a compensation measure, a manually curated tabular dataset extracted from the table of contents was introduced as a hierarchical reference for certain headings. It was initially aimed at including the hierarchical information contained in the predictions of the Layout Model in a further evolutionary revision of the python code. However, before this revision could be initiated, members of the project team of the Schematismus Project had produced a very similar reconstruction of the hierarchical network and had implemented the same as a graph database. The reconstructions of the project team considered not only sequential positioning of elements, but also the predicted layout category and a manually curated reference to the table of contents.

[124] THE PANDAS DEVELOPMENT TEAM, pandas-dev/pandas: Pandas 2024. Online reference: pandas.pydata.org/, (last accessed: 02.02.2025).

[125] Digitized by the ÖNB and publicly accessible under: alex.onb.ac.at/cgi-content/alex?aid=shb&datum=1878&page=196, (last accessed: 06.01.2025).

Protiwensky Wilhelm, Bes. d. Kgs. Med. u. d. pers. S. u. L. O. V.
Kállay Béla v. Nagy-Kálló, Bes. d. Kgs. Med. (prov.).

Hof-Fourier-Ansager.

Latty Jacob, Bes. d. Kgs. Med.

Hof-Stabs-Adjutant.

Klinger Johann, Oberlt., wie Seite 27.

Hof-Stabs-Feldwebel.

Troszt Anton, Bes. d. slb. T. M. l., d. Kgs. Med. u. d. r. G. Kr. IV.
Hauffe August, Bes. d Kgs. Med., d. r. gld. M. (a. St. Bd.) u. d. slb. M. d. pers. S. u. L. O.

Administrationen.

K. K. Hof-Bibliothek.

(Hofburg, Josephsplatz 1.)

Vorstand.

Birk Ernst, R. d. L. O. u. d. Fr. J. O., Ph. Dr., Mitgl. der Ak. d. Wiss. in Wien, ausw. Mitgl. der ung. Ak. d. Wiss., Corresp. d. Centr. Cmsn. f. Kunst- u. histor. Denkmale, E. Mitgl. des Ferdinandeums in Innsbruck, der histor. Ver. f. Steiermark und Kärnten u. der histor. Sect. des Ver. f. Landeskunde in Brünn, Mitgl. des histor. Ver. von u. für Ober-Bayern, des Gel. Aussch. des germanischen Mus. zu Nürnberg u. des Alterthums-Vereins in Wien, corr. Mitgl. der Ak. d. Wiss. in München, des Ver. f. Landeskunde Siebenbürgens u. der oberlausitzischen Ges. der Wiss. zu Görlitz, k. k. wirkl. H. R.

Custoden.

Pachler Faust, J.Dr.
Raab Ferdinand, R. d. Fr. J. O.
Haupt Joseph, corr. Mitgl. der Ak. d. Wiss. in Wien.

Scriptoren.

Hartl Wenzel, Bes. d. Kgs. Med.
Müller Friedrich, Ph.Dr., ordentl. Prof. d. Sanskrit u. der allgem. Sprachwissenschaft an der Wiener Univ., Mitgl. der Ak. d. Wiss. in Wien, ord. ausw. Mitgl. d. Ak. d. Wiss. in München, Mitgl. u. Vice-Präsdt. der anthropologischen Ges. in Wien, ordentl. Mitgl. der kais. Ges. der Naturforscher in Moskau, der anthropologischen Ges. in Paris u. des koninklijk Instituut voor de taal-, landen volkenkunde van Nederlandsch Indie in Haag, E. Mitgl. der Société philologique (langues Anaryennes) in Paris und der kön. italienischen Ges. für Anthropologie u. Ethnologie in Florenz.
Wöber Franz.
Majr Joseph.
Göldlin v. Tiefenau Alfred, Ph.Dr.
Kaltenleitner Joseph, Ph.Dr.

Amanuenses.

Göttmann Carl.
Cammerloher Moriz.
Mančik Ferdinand.

K. K. naturhistorisches Hof-Museum.

Intendant.

Hochstetter Ferdinand, v., R. d. eis. K. III., Cmdr. d. bras. R. O. u. d. belg. L. O., R. d. würt. Kr. O., Bes. d. ottom. M. O. III., Ph.Dr., Mitgl. der Ak. d. Wiss., Mitgl. der wiss. Realschul-Prüf. Cmsn., Präsdt. der geographischen Ges. in Wien, Prof. der Mineralogie u. Geologie an der technischen Hochschule in Wien, H.R.

Assistent.

Heger Ferdinand.

K. K. zoologisches Hof-Cabinet.

(Hofburg 1.)

Director.

Steindachner Franz, R. d. Fr. J. O., Ph.Dr., wirkl. Mitgl. der Ak. d. Wiss. in Wien, corr. ausw. Mitgl. der kön. Ak. d. Wiss. zu Lissabon, der zoolog. Ges. in London u. der naturwiss. Ak. zu San Francisco in Californien, Mitgl. der Ak. Leop. Car., der Société d'Acclimatation zu Paris, der Société zoologique de France, des siebenbürgischen Ver. f. Naturwiss. in Hermannstadt, der zoologisch-botanischen Ges., der k. k. geogr. Ges. u. des ornithologischen Vereines in Wien, Mitgl. des Vereines für Landeskunde von N. Oe., E. Mitgl. der Literary u. Philosophical Society of Liverpool etc.

Custoden.

Pelzeln August, v., E. Mitgl. der British Ornithologist's Union, Präs. d. ornithologischen Vereins in Wien, Mitgl. der zoologisch-botanischen Ges. u. der deutschen ornithologischen Ges. zu Berlin.
Rogenhofer Alois Friedrich, 1. Secr. der zoologisch-botanischen Ges. in Wien, Mitgl. der Ak. Leop. Car., des Ver. für Landeskunde von Nieder-Oesterreich, der entomologischen Ver. zu Berlin, Brüssel, Florenz, Stettin u. St. Petersburg, des Ver. der Naturfreunde zu Reichenberg, des deutsch-österr. Alpen-Ver., der naturwiss. Ve. zu Graz u. Wiesbaden, des ornithologischen Ver. in Wien, Corr. der geologischen Reichsanstalt.
Brauer Friedrich, M.Dr., a. o. Prof. an der Univ., Docent an der Hochschule für Bodencultur, Bes. d. gld. Med. f. Kst. u. Wiss., Ausschuss-rath d. zoologisch-botanischen Ges. in Wien, Mitgl. mehrerer gel. Ges. des In- u. Auslandes.
Marenzeller Emil, Edl. v., MDr., emerit. Assistent der zoolog. Lehrkanzel an der Univ. zu Wien, Secr. der zoologisch-botanischen Ges. in Wien.

Assistenten.

Koelbel Carl, Mitgl. mehrerer gelehrten Gesellschaften.
Krauss Hermann, M. & Chir. Dr., Mtgl. der zoologisch-botanischen Ges. im Wien.

Aufseher.

Mann Joseph, Bes. der slb. M. mit dem Lorbeerkranze für Kst. u. Wiss., des kön. Mus. in Florenz, Mitgl der zoologisch-botanischen Ges. im Wien, des naturforschenden Ver. „Lotos" in Prag, der naturforschenden Ges. in Görlitz, des pomologischen Ver. in Zittau, des entomologischen Ver. zu Stettin, corr. Mitgl. der Ges. f. specielle Naturgeschichte in Dresden.
Werner Theodor, Bes. d. Kgs. Med.

Präparatoren.

Tonnebaum Ferdinand, Bes. d. Kgs. Med.
Zelebor Rudolph.

K. K. mineralogisches Hof-Cabinet.

(Hofburg 1.)

Director.

(Unbesetzt.)

Provisorisch mit der Leitung betraut:
Hochstetter Ferdinand, v., H. R., wie oben.

Custoden.

Fuchs Theodor.
Březina Aristides, Ph.Dr., Privat-Docent für Mineralogie an der Univ. in Wien.

Assistent.

Berwerth Friedrich, Ph.Dr.

Kanzlist.

Engelmayer Anton.

Aufseher.

Auinger Mathias.
Brattina Franz.

Figure 19: Page 18 of the Schematismus of 1878.

It was therefore decided to not further evolve the python code shown in Annex 1 but rather compare and evaluate the completeness of both reconstruction efforts – the one produced by the sequential logic presented herein and the other produced by the project team. This comparative evaluation will be the main concern of chapter 5.3. Before that, the practical implementation of the sequence-based reconstruction effort will be explained in detail.

5.2.1 Handling of 'H6' Heading Classes

The application of the logically deducted rules to decipher sequential positioning is impeded by headings of the 'H6' category. On page 18 of the 1878 Schematismus there are four distinct 'H6'-elements. Respectively they add an address to an institutional entity (represented by the preceding heading) in three of the cases and explain that the position of "Director" of the "K. K. mineralogisches Hof-Cabinet" (mineralogical cabinet) is vacant ("Unbesetzt.") in the year 1878 in the fourth case. These two different types of information have different implications for the hierarchical structure. While the address can be considered an extension of an existing heading, the interim placeholder for an absent individual within the organization might structurally be viewed as a paragraph.

In order to enhance comprehension of the various types and the frequency of the 'H6' class in the dataset, a separate data sample was extracted from the existing data, containing all instances of 'H6'-headings within the initial 189 pages (i.e., the entire court section) and the 30 pages relating to the regional administration of Styria. The data subset yielded the moderate amount of 147 'H6'-instances within the court section and a remarkable 150 'H6'-instances within the 30 pages of the Styrian administration. For the purpose of preliminary classification, the 'H6' elements in the sample were assigned to one of six types respectively:

- address: The address-type always follows directly after a heading, thereby appending supplementary information about the physical address of the respective administrative unit.
- pers_equiv: The pers_equiv-type is used as a placeholder for an individual officer. Most of the time it takes form of "(Unbesetzt.)", which means that the respective position is vacant in the respective year.
- info: The info-type is structurally similar to the address-type, as it ascribes additional information to the preceding heading class.

- none: The none-types are very rare cases, where there is no additional information contained in the CSV-data regarding the respective ‘H6’ element.
- mistake: The mistake-type is assigned, when the ‘H6’ class is wrongly assigned.
- categorization: The categorization-type is used to label those ‘H6’ headings that actually fulfill a heading purpose.

The six categories proved sufficient to appropriately handle the approximately 300 cases that appeared in the subsample. There might be methods to automatically categorize the ‘H6’-headings, but considering the limited number of instances, it was decided to assign the type-information manually. For this purpose, a column entitled “H6_type” was added to the data subset. Subsequently, the subset was recombined with the original dataframe.

```
# handle case pers_equiv: the class is set to Paragraph
court_df.loc[court_df['H6_type'] == "pers_equiv', "class"] = "Paragraph"

# handle cases address and info by creating new columns that contain the respective text
court_df.loc[court_df['H6_type'] == "address", "address"] = court_df.loc[court_df[
                'H6_type'] == "address", "pred_text"]
court_df.loc[court_df['H6_type'] == "info", "info"] = court_df.loc[court_df[
                'H6_type'] == "info", "pred_text"]

# these new columns need to move a row up, as it is additional information
# to the preceding element
court_df["address"] = court_df["address"].tolist()[1:] + court_df[
                "address"].tolist()[:1]
court_df["info"] = court_df["info"].tolist()[1:] + court_df["info"].tolist()[:1]

# drop the H6 rows that are not needed anymore
court_df = court_df.loc[court_df['H6_type'].isin(["address", "info", "none"]) == False]
```

Figure 20: Code snippet showing how the different ‘H6’ categorizations are handled.

In order to reflect the implications for reconstructing hierarchical relationships, each ‘H6’-type was handled differently during the merging operation. ‘H6’-elements of the ‘pers_equiv’ type were re-designated from ‘H6’ to ‘Paragraph’ classes. The textual content of the ‘address’ and ‘info’ types were used to augment the preceding rows with supplementary information. The respective lines were removed from the data, and the textual value assigned to newly introduced columns in the previous row. This approach ensures the preservation of information for future use. The ‘none’ types do not possess any additional

value and are thus eliminated. In contrast, the 'mistake' values were so infrequent that they were addressed manually. The remaining 'categorization' types keep their heading function.

5.2.2 Implementing a Sequential Logic

Having established the manner by which 'H6'-headings are to be handled, the focus can return to implementing the logical rules by which the hierarchical relationships of the Schematismus can be reconstructed. A summary of the previously deducted rules to decipher sequential positioning reads as follows: Successive paragraphs are of the same organizational level and are structurally subordinate to the preceding heading. Successive headings on the other hand are of different organizational levels, where the first of the sequence occupies the highest and the last one occupies the lowest rank. The order decreases steadily. Headings that follow after paragraphs belong to the same organizational level as the last preceding heading.

The Python code under consideration constitutes an attempt to implement those rules on the dataset. The implementation is facilitated by hierarchy values that are assigned to each row of the CSV-tables. Adopting this approach the lowest numerical rating "0" is allocated to each 'Paragraph'. Each heading immediately preceding a 'Paragraph' is assigned a rating of "1". For each heading that follows in a direct reverse sequence, the rating is incrementally increased by 1.

In this manner, the hierarchy value is initially reconstructed from the bottom up. Subsequently, hierarchical superiority is determined through a systematic comparison of the assigned hierarchy ratings top down. The zone of influence of a heading with a given hierarchical level extends until a subsequent heading of the same or higher level is encountered. Figure 21 illustrates the application of this logic to a section of paragraphs and headings. The column adjacent to the predicted classes

Paragraph	0
Paragraph	0
H1	2
H2	1
BigPrargraph	0
H2	2
H5	1
Paragraph	0
Paragraph	0
Paragraph	0
H5	1
Paragraph	0
Paragraph	0
H5	1
Paragraph	0
H5	1
Paragraph	0
Paragraph	0
Paragraph	0
H2	1
BigPrargraph	0
H5	2
H5	1
Paragraph	0
Paragraph	0

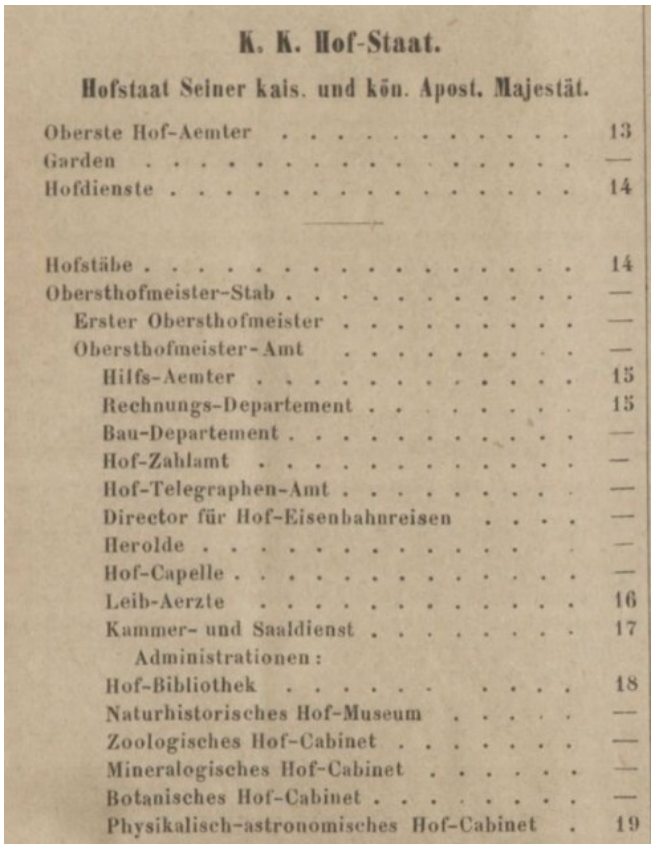
Figure 21: The table to illustrate the rule of sequence.

contains the numerical ratings that the logic would assign to the respective elements. The color code located on the right side of the table delineates the zone of influence of headings. The yellow sections of the table indicate elements that belong to headings with the rating class “1”. The blue section serves the same function for elements pertaining to headings with the rating class “2”.

The implementation of this logic aims at resolving one of the primary challenges of the present reconstruction effort, namely determining the extent of the zone of influence for each heading provided in the CSV data. As those zones can only be correctly identified, if the hierarchical difference between consecutive headings can always be correctly determined, this approach is expected to be flawed. The lower portion of the table in Figure 21 is illustrative of such an irregularity. Due to the logic of sequential positioning and the disregarding of predicted categories, the final ‘H2’ heading gets a rating of “1”, while the succeeding ‘H5’ heading gets a rating of “2”.

5.2.3 Introducing Auxiliary Information

Due to the nature of the provided information, which consists of abstracted hierarchical data and a consolidation of multiple heading styles into five distinct classes, it was obvious from the outset that reproducing the organizational tree in its entirety was not feasible. Consequently, the idea arose to compensate for the missing information through a manually curated tabular dataset extracted from the table of contents. This dataset served as an external knowledge base regarding the hierarchy of the elements listed in this register. The table of contents of the 1878-Schematismus is fourteen pages long and contains some hierarchical indications for the upper levels of the organizational structure. As illustrated by Figure 22, the hierarchical information is visually conveyed through the use of line indentations.


K. K. Hof-Staat.

Hofstaat Seiner kais. und kön. Apost. Majestät.

Figure 22: Screenshot from the table of contents of the Schematismus.

The extract shown in Figure 22 can also be used to illustrate why a manual curation of the dataset is advisable. Regarding the entry "Administrationen", additional context knowledge is required to correctly interpret that it is in fact not a sub-heading of "Kammer- und Saaldienst" and that the succeeding headings are themselves sub-headings of the "Administrationen". Furthermore, the exact spelling of the entries in the table of contents may not always correspond to the spelling used on the pages referenced by the register. For instance, the court of Emperor Franz Joseph is referred to as "Hofstaat Seiner kais. und kön. Apost. Majestät." in the table of contents, yet on page 13 of the Schematismus the same appears as "Hofstaat Seiner Oesterreichisch-kaiserlichen und königlich-Apostolischen Majestät." In certain instances, a solitary entry in the register may be associated with multiple headings in the primary section. All these factors impede an automated extraction approach.

```python
def redo_hierarchies(df):
    section_df = df

    # If the element has a hierarchy level assigned from the TOC, this level is copied to the
    # "institutional_hierarchy" column
    section_df.loc[section_df.TOC_hierarchy.isin(range(21)),
                   "institutional_hierarchy"] = section_df.loc[
                       section_df.TOC_hierarchy.isin(range(21)), "TOC_hierarchy"]

    # I only work on the hierarchy column and therefore convert it to a list
    re_hierarchy = section_df["institutional_hierarchy"].values.tolist()

    # I determine the validated values in the hierarchy column. I define a lower cut-off at 9.
    validated_vals = [x for x in set(re_hierarchy) if x>=9]
    # I determine the indexes of the validated values in the list
    validated_indexes = [i for i,x in enumerate(re_hierarchy) if x in validated_vals]
    # The indexes of the validated values define range boundaries between which the hierarchy
    # values need to be adjusted
    ranges = [0] + validated_indexes + [len(re_hierarchy)]
    # I determine the highest hierarchy value of the non-validated values
    non_val_vals = [num for num in re_hierarchy if num not in validated_vals]
    max_val = max(non_val_vals)

    # for each range between two validated values, I adjust the non-validated values
    # values in ranges need always be lower than any validated value
    diff = 0
    for i in range(len(validated_indexes)+1):
        for u in range(ranges[i], ranges[i+1]):
            # for the first value in the range, I set increment as the difference between the
            # itself and the highest non-validated value
            # such, each that the values in the range decrease by one starting from the level of
            # the first value in the range
            if re_hierarchy[u] in validated_vals:
                diff = re_hierarchy[u] - max_val
            else:
                re_hierarchy[u] += diff - 1
```

Figure 23: An extract from the programming notebook, showing python code used to integrate curated hierarchy values from the table of contents into the data set.

Considering the potential benefit of a reliable reference table for the reconstruction efforts, the manual curation of such a dataset appeared to be a worthwhile endeavor. For

the purpose of this master's thesis, hierarchical information from the table of contents was manually extracted for the section of the court of Emperor Franz Joseph and the regional administration of Styria.

It has been indicated that the python code underwent iterative modifications to align with the layout properties of other pages. In essence, the observed adaptations were predominantly due to an escalation in complexity and the presence of more hierarchical levels within a single page. The need for a reference by which to calibrate the hierarchical levels emerged as soon as attempts were made to apply the code longer sections across multiple pages.

The practical implementation of the curated hierarchy levels saw the creation of continuous dataframes, combining all CSV pages of the court and the regional administration of Styria, respectively. A new column, designated "TOC_hierarchy" (TOC stands for "table of contents"), was incorporated, along with number values that indicate the hierarchical order of all elements that were referenced in the table of contents. Left-aligned values in the table of contents were assigned the value 20 and with every indentation the value was reduced by 1. Thus, the lowest hierarchy value in the curated TOC-dataset was 15, while the highest hierarchy value established through the logic of sequential positioning – i.e. summing up successive headings in reverse order – was 4. This allowed to define a cut-off value to securely identify the curated hierarchy values. These calibrated values were kept constant while all hierarchy values of headings between two calibrated headings were readjusted. The readjustment entailed closing the gap of the hierarchy value between the calibrated and the first of the succeeding headings, such that the latter occupied the next lower hierarchy level. The remaining non-calibrated headings were handled accordingly, in order to maintain their respective hierarchical relationship. The code snippet in Figure 23 depicts the respective lines of python code.

5.2.4 Disambiguating Entities

Having established and implemented rules, by which the sequence of layout elements can be used to deduct preliminary hierarchical relationships, this short section shall briefly address the challenge of reliably identifying and disambiguating the three core entity types 'Persons', 'Positions', and 'Institutions'. While the layout classification provides a robust foundation for identifying 'Persons' through 'Paragraph' elements, the distinction between 'Positions' and 'Institutions' within the heading elements necessitates further analytical steps. The 'Position' entity plays a pivotal role in the data model, functioning as a

conceptual bridge between individual civil servants and the overarching institutional framework. However, this entity type is not explicitly identified through the layout classification system. In order to address the aforementioned limitation, a heuristic approach was developed that utilizes sequential relationships between layout elements to differentiate between headings representing institutional units and those representing positions within said units.

```
# all headings carry some information on organisational structure or hierarchy

# I introduce a new column "is_orga" that is true if the class is H0, H1, H2, H3, H4, H5
or H6

section_df["is_orga"] = np.where((section_df['class'] == "H0") |

                                 (section_df['class'] == "H1") |

                                 (section_df['class'] == "H2") |

                                 (section_df['class'] == "H3") |

                                 (section_df['class'] == "H4") |

                                 (section_df['class'] == "H5") |

                                 (section_df['class'] == "H6"), True, False)


# to determine whether a heading is a function or an institution, I need to look at the

# subsequent element. If the subsequent element is a paragraph, the heading is a function.

section_df["is_function"] = section_df["class"]

    for i in range(len(section_df)):

        if section_df.loc[i, "is_function"] in type_heading and section_df.loc[

            i+1, "is_function"] in type_paragraph:

            section_df.loc[i, "is_function"] = True

        else:

            section_df.loc[i, "is_function"] = False

# if a heading is a function, it is not an institution

section_df["is_institution"] = (section_df["is_function"] - section_df[

    "is_orga"]).astype("bool")
```

Figure 24: Code snippet showing simple boolean indexing to disambiguate between institutions, positions and persons.

This solution entails the creation of Boolean index columns within the dataset. The columns “is_person”, “is_orga”, and “is_function” encode the disambiguation between the three entity types. The term “is_function” in the code is equivalent to the ‘Position’ entity in the data model, while “is_orga” is equivalent to ‘Institution’. The “is_person” column is applicable to all ‘Paragraph’ types, while the “is_orga” column is applicable to all headings. The “is_function” column is labeled “True” for all headings that immediately precede a paragraph. The relevant section of the code can be observed in Figure 24.

As discussed in chapter 4.4, the ‘Position’ entity will often be presented as headings but can also be encoded as appended information within paragraphs. As a result the presented solution has to be viewed as preliminary and the conclusive assignment of entities postponed until named entity recognition steps are performed later in the Schematismus Project.

The presented rule-based approach, which implements a logic to translate sequential positioning of headings into hierarchical relationships, was insofar completed, as it can be applied and visualized for multi-page sections of the Schematismus. Otherwise, functional development was stopped after the incorporation of the curated table of contents. The cessation of further development was due to the mentioned parallel reconstruction effort on behalf of the team of the Schematismus Project. At the same time, visualization options for the presented rule-based approach were implemented, to enable a comparison of results with the efforts of the Schematismus Project team. While the comparison of the two experiments is performed in the next chapter, the entire code base of the reconstruction approach presented in this chapter can be found in Annex 1: Commented Code Repository.

5.3 Comparative Evaluation of Reconstruction Efforts

The reconstruction approach presented in the previous chapter attempts to derive hierarchical relationships between entities based on their sequential positioning within the structured output of the Layout Model. Additionally, it uses a manually curated table of contents register to introduce fixed hierarchy levels for certain entities. It was mentioned in chapter 5.2 that a similar reconstruction effort was performed by members of the project team of the Schematismus Project partly in parallel. The hierarchical network, which resulted from this approach draws also on the structured output from the Layout Model and a curated table of conted – although a different one – but it additionally takes into account the predicted heading categories. Both efforts can be considered experimental, and their

resulting hierarchical networks need to be evaluated regarding their accuracy. This evaluation will be performed exemplary, using selected sections from the Schematismus of 1878. In this way the outputs of the two experimental reconstruction efforts can also be directly compared.

The Schematismus Project team implemented their reconstructed network as a knowledge graph in Neo4j. The knowledge graph is comprised of two types of edges, namely “features on” and “is part of”. The “features on” edge type establishes a connection between each component of a page and its respective page number. The edge type “is part of” denotes a connection from an element to its immediate superior within the designated hierarchy. The approach is applied to the data of the entire 1878 edition and is hosted on a Neo4j server, making it accessible for a comparative analysis.

The following three examples have been selected from different sections of the Schematismus and include various frequent peculiarities of the layout. The subsequent discussion will explore the differences between the resulting hierarchical structures of the two experimental approaches. The screenshots from the Neo4j visualization are characterized by a color-coding system. Provided that a certain degree of sorting effort is invested, the Neo4j graphs can be effectively organized. The supplementary links that extend away from each node and terminate beyond the confines of the screenshot are attributable to the fact that each node is associated with the page number to which it corresponds.

To visualize the respective results from the approach presented in chapter 5.2, the Networx-library[126] for Python was used. Regrettably, the visualizations are not easily incorporated as screenshots, because the library produces PNG files of an excessively wide dimension.

[126] NETWORKX DEVELOPERS, NetworkX 2024. Online reference: networkx.org/, (last accessed: 02.02.2025).

5.3.1 Example: Page 18

The first example uses a short section from page 18 (Figure 19). The bottom half of the right column depicts the members of the "K. K. mineralogische Hof-Cabinet". Evidently, the position of "Director" was vacant in 1878. As can be seen in Figure 25, the vacancy place-holder ("Unbesetzt") is recognized as 'H6' class, in the same way as the address "Hofburg 1" two lines earlier. The heading indicating a provisional head of the department ("Provisorisch mit der Leitung betraut") is categorized as 'H5'. A comparison with the print reveals that the style of this heading differs in comparison with other surrounding 'H5' headings.

```
0438113,H2,0.9413874745368958,"K. K. mineralogisches Hd-

1802,H6,0.8772698640823364,(Hofburg 1.)
777268,H5,0.8849978446960449,Director.
7322303,H6,0.9126347303390504,(Unbesetzt.)
1090686,H5,0.8803191781044006,"Provisorisch mit der Leiting

1072304,Paragraph,0.975366234779358,"Hochstetter Ferdinasd, v., H. R.,

745864,H5,0.9026711583137512,Custoden.
5196384,Paragraph,0.9638269543647766,Fuchs Theodor.
5859682,Paragraph,0.976055920124054,"Březina Aristides, Fh. Dr, Prwat-
```

Figure 25: Screenshot from the CSV-data extracted from page 18 of the Schematismus.

The rule-based approach that was introduced in the previous chapter utilizes a manually curated 'H6' dictionary. The outputs of this approach are visualized in Figure 26. The bespoke handling of 'H6' classes places the vacancy placeholder on the hierarchical level of a person and does not treat the address as an entity. The address information is assigned to the Institution-entity of the mineralogical department as attribute and thus not visible in the visualization.

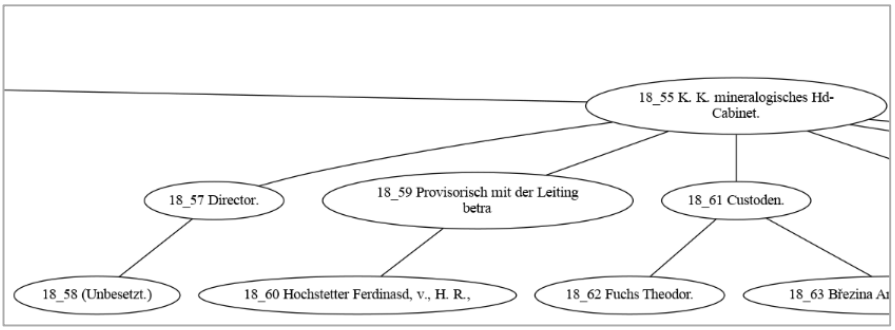


Figure 26: Visualization of the links created for the bottom section of page 18 by the reconstruction heuristic, which is based on the rule of sequence.

The comparative solution that has been visualized in Neo4j is presented in Figure 27 below. The organizational structure is slightly different. The heading classes are designated by a pink color, whilst the paragraphs are indicated by a green color. The predicted 'H6' classes ("Unbesetzt" and "Hofburg 1.") are both modeled as entities. In this solution the provisionary director ("Provisorisch mit der Leitung betraut.") is modeled as a sub-node to "Director", which seems a slightly more elegant solution than introducing the same as a separate institutional entity belonging to the mineral department (compare Figure 26). It is, however, evident that both approaches yield solutions that can be considered correct for this short segment.

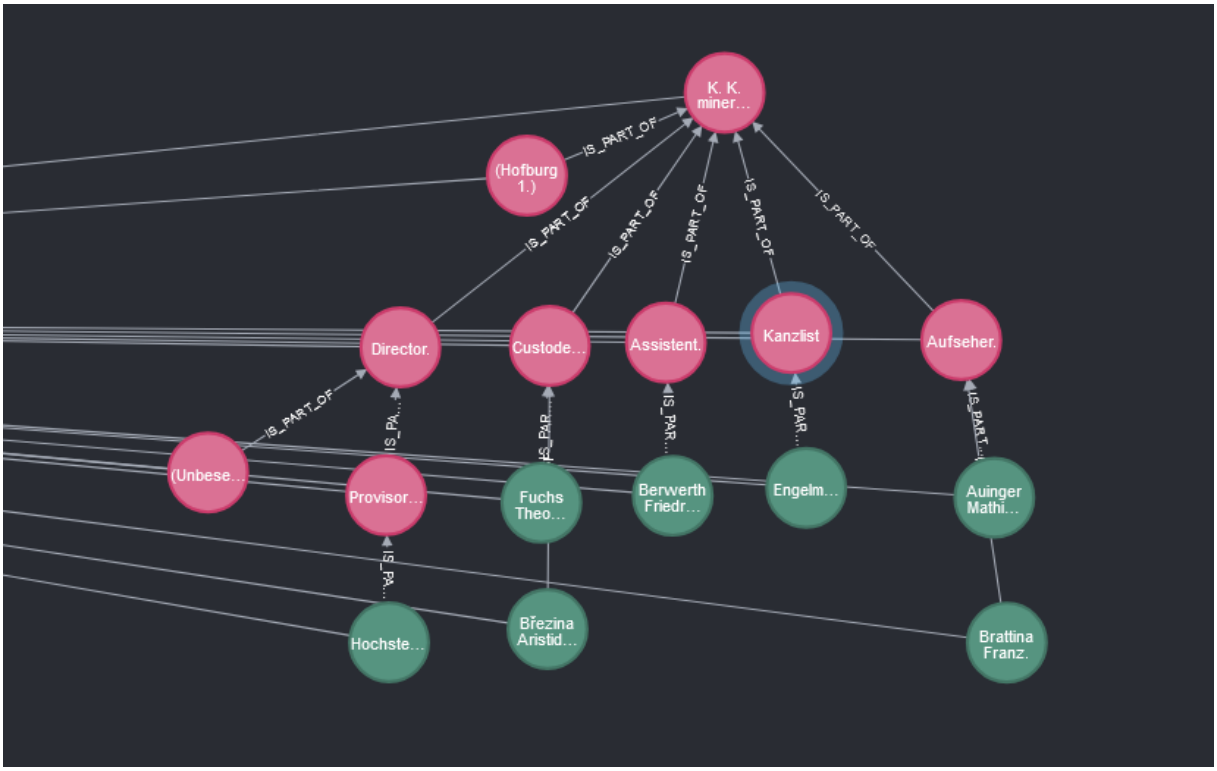
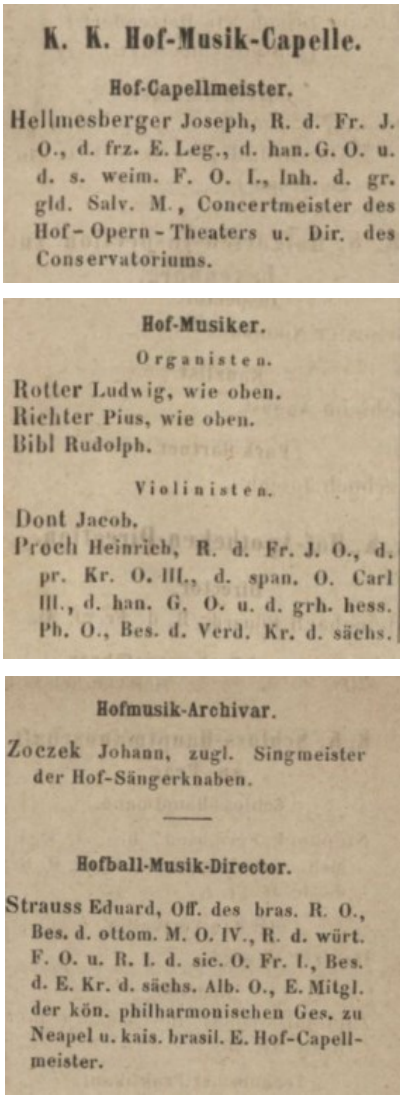


Figure 27: Screenshot from the Neo4j visualization of the bottom section of page 18.

5.3.2 Example: K. K. Hof-Musik-Capelle

The “K. K. Hof-Musik-Capelle” (Court Orchestra) on page 21 of the 1878 Schematismus will serve as an example that the lack of disambiguation of heading styles within the lower ranks of the organizational structure poses a problem for reconstruction efforts. The “K. K. Hof-Musik-Capelle” is referenced in the table of contents and thus possesses a guaranteed hierarchical value. The orchestra as an institutional entity is divided into departments such as singers, musicians and among others the court archivist. All its sub-headings are predicted to be ‘H5’ classes. Consequently, there is no discrimination between, for example the heading “Hof-Musiker” and the various sections of instrumentalists that pertain to it.


K. K. Hof-Musik-Capelle.

Hof-Capellmeister.

Hellmesberger Joseph, R. d. Fr. J. O., d. frz. E. Leg., d. han. G. O. u. d. s. weim. F. O. I., Inh. d. gr. gld. Salv. M., Concertmeister des Hof-Opern-Theaters u. Dir. des Conservatoriums.

Hof-Musiker.

Organisten.

Rotter Ludwig, wie oben.
Richter Pius, wie oben.
Bibl Rudolph.

Violinisten.

Dont Jacob.
Proch Heinrich, R. d. Fr. J. O., d. pr. Kr. O. III., d. span. O. Carl III., d. han. G. O. u. d. grh. hess. Ph. O., Bes. d. Verd. Kr. d. sächs.

Hofmusik-Archivar.

Zoczek Johann, zugl. Singmeister der Hof-Sängerknaben.

Hofball-Musik-Director.

Strauss Eduard, Off. des bras. R. O., Bes. d. ottom. M. O. IV., R. d. würt. F. O. u. R. I. d. sic. O. Fr. I., Bes. d. E. Kr. d. sächs. Alb. O., E. Mitgl. der kön. philharmonischen Ges. zu Neapel u. kais. brasil. E. Hof-Capellmeister.

Figure 28: Screenshots from Schematismus page 21 showing the “K.K. Hof-Musik-Capelle”.

Since “Hof-Musiker” and “Organisten” are directly sequential headings, the rule-based approach presented in this thesis has the capacity to perform the step down the organizational hierarchy from the overarching group of musicians to the several instruments of the orchestra. However, the data does not provide any indication that the archivist (“Hofmusik-Archivar”) should not be considered part of the musicians and should instead link back to the institutional entity of the orchestra. As is apparent in Figure 29, although not fully discernible, the nodes labeled “Pauker”, “Hofmusik-Archivar” and “Hofball-Musik-Director” are all connected to the same node: “Hof-Musiker”.

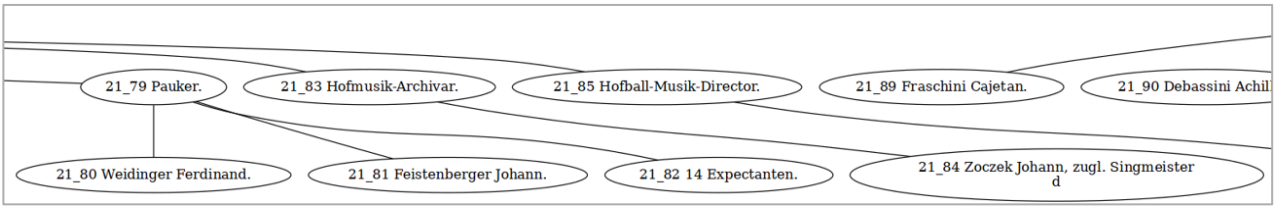


Figure 29: Visualization of the links created for the “K.K. Hof-Musik-Capelle” by the reconstruction heuristic, which is based on the rule of sequence.

The Neo4j graph shows a differently linked organizational tree. It is evident that the institutional entity “Hof-Musiker” is linked back to the orchestra. The initial group of instrumentalists, the “Organisten”, link back to the “Hof-Musiker”. This outcome is in accordance with the anticipated results. However, continuing in reading order, the remaining instrumentalists are linked directly to the overarching institutional entity “K. K. Hof-Musik-Capelle”, resulting in the formation of multiple erroneous connections. The situation is illustrated in Figure 30. Conversely, the “Hofmusik-Archivar” is correctly linked directly to the orchestra as the superior administrative unit.

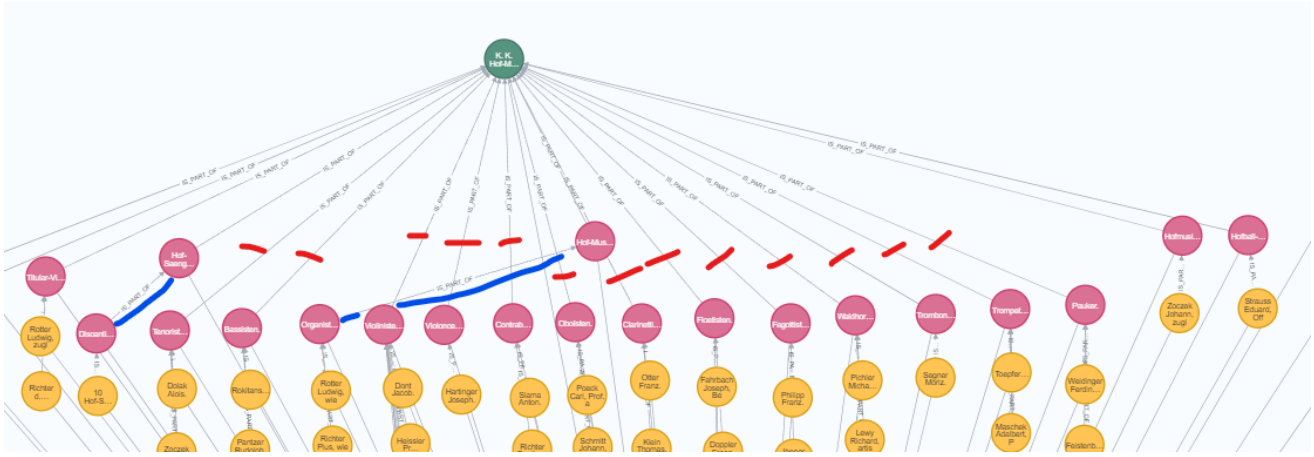


Figure 30: Screenshot from the Neo4j visualization of the “K.K. Hof-Musik-Capelle”.

This example demonstrates that both heuristics are flawed and are unable to produce entirely correct solutions based on the given information. The results and errors obtained in this case vary, depending on whether the predicted hierarchy class or a rule of sequence is prioritized.

5.3.3 Example: Curly Brackets

The correct interpretation of curly brackets has been explained in section 4.3.3. The bracket serves to visually delineate the area of influence exerted by the text block to its right. As illustrated in Figure 31, the bracket is utilized to assign the ‘Position’ entity designated “Secretäre” to the two ‘Person’ entities to its left. The “Secretäre” text block functions as a heading for these two paragraphs and is correctly recognized as such (‘H5’) by the Layout Model. As is common for sections with curly brackets, a respective heading at the same


19. Windisch-Graz.

(Ger. Bez. Windisch-Graz, Schönstein, Mahrenberg.)

(Unbesetzt.)

Landes-Beamte.

Wanggo Emanuel, } Secretäre.
Gstettenhofer Jos., Dr., }
Schaffernak Alois, L. Buchhalter.
Koch Anton, L. Ob. Einnehmer.
Bartl Ernest, L. Bau-Dir.
Mitransky Fr., L. Gebäude-Inspector.
Podgorschegg Ed., L. Hilfsämter-Dir.

Figure 31: Curly bracket on page 329 of the Schematismus.

hierarchical level is missing for the subsequent paragraphs. These individuals carry the position information as appended text within their paragraph.

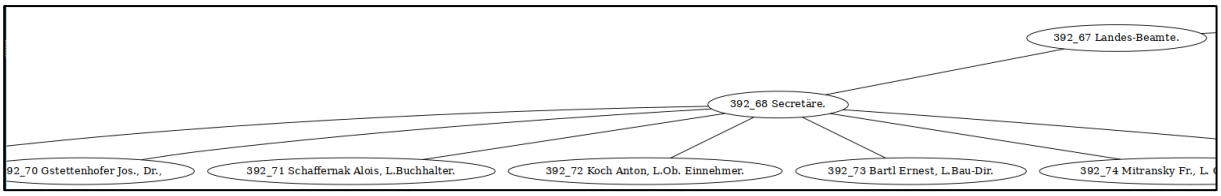


Figure 32: Visualization of the links created for the curly bracket section on page 329 by the reconstruction heuristic, which is based on the rule of sequence.

The visualization produced by the rule-based approach introduced in this thesis is shown in Figure 32. The line "Secretäre" is identified as a 'H5' heading by the Layout Model and is positioned in front of the relevant paragraphs by the resorting algorithm. Consequently, the positional logic links all following successive paragraphs to the Position entity "Secretäre".

Figure 33 presents the same section but extracted from the graph database of the Schematismus Project team. It appears that they have integrated the bounding box data of the curly container to ascertain the extent of the curly container. The separation of the two paragraphs affected from the curly bracket and the subsequent entries is facilitated by this information. In light of the current unavailability of semantic information on the textual level, this solution represents the most viable option at present.

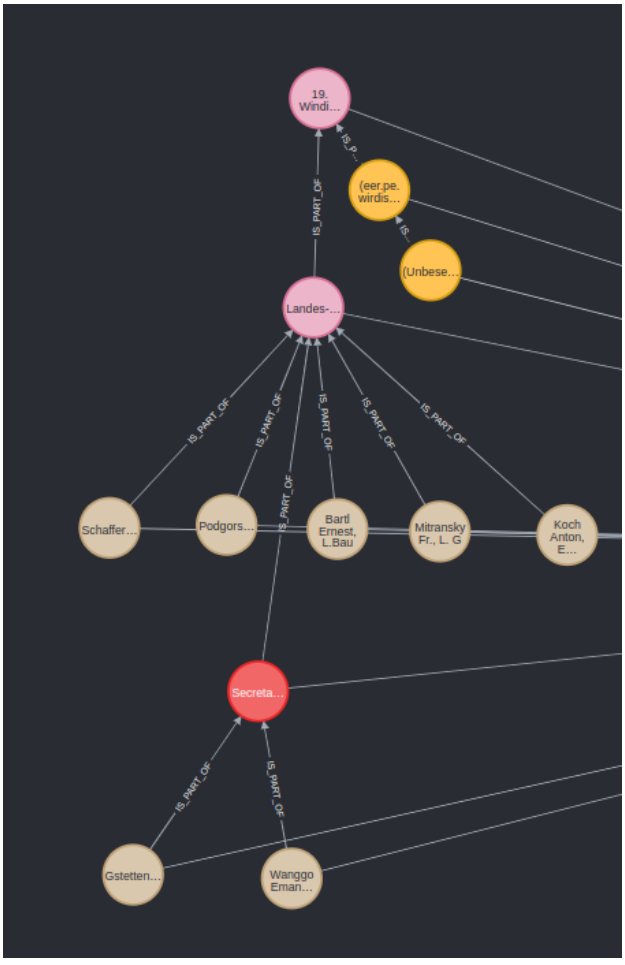


Figure 33: Screenshot from the Neo4j visualization of the curly bracket section on page 329.

However, it is imperative to draw attention to another noticeable peculiarity observable in Figure 33. The heading "Landesbeamte" is linked to the heading "19. Windisch-Graz". A comparison with Figure 31 reveals the inaccuracy of this assertion. The heading "Landesbeamte"

indicates the beginning of a new section. “19 Windisch-Graz” constitutes merely a regional subdivision of the preceding section. The erroneous linkage can be attributed to two primary factors: Firstly, the distinction on the layout level was inadequate. The two headings in question were both labeled as instances of the ‘H5’ class. Secondly, the reconstruction logic is erroneous due to an absent ‘Paragraph’. The respective Position entity in Windisch-Graz was unoccupied in the year 1878, and the respective placeholder element was correctly classified as ‘H6’. The absence of a ‘Paragraph’ results in the assumption of a continuous cascade of hierarchies. The same inaccuracy does not occur using the rule-based approach introduced in chapter 5.1 because the ‘H6’ classes are addressed through a manually curated dictionary, which maps the “Unbesetzt” placeholder to the ‘Paragraph’ level.

Concluding the evaluation of the two experimental hierarchy reconstruction approaches, it is evident that none of them can reproduce the organizational structure of the printed Schematismus with satisfying precision based on the currently available information. Out of three short sections, which were of course selected due to their inherent complexity, only one was reconstructed at a satisfactory level.

5.4 Summary of Reconstruction Efforts

In the preceding sections, a comparison was made between two approaches to the reconstruction of organizational hierarchies from extracted layout data. It is evident that both of these attempts have been unsuccessful in deciphering the nested structure of the Schematismus.

The first approach, explained in detail in chapter 5.2, tries to establish a rule purely based on the positional sequence of elements in reading order. The examples demonstrate that this approach is unsuccessful in instances where an organizational element possesses multiple child elements, which themselves may or may not encompass further child elements. The deviation from the correct structure manifests then at the transition from an organizational element with child to one without. The table shown in Figure 34 displays the institutions (headings) of the “K. K. Hof-Musik-Capelle” in the first column (see Figure 28 for the printed version). The remaining columns hold the predicted layout classes, the hierarchy rating assigned by the sequence-based reconstruction logic and the same color codes as in Figure 21. Green highlights in the first column indicate first degree sub-levels of the court orchestra. Of those headings only “Hof-Sänger” and “Hof-Musiker” possess an additional tier of sub-headings (violet). The sequence-based logic can successfully identify

the commencement of a sub-section but fails to identify the next same-level heading that doesn't exhibit a sub-heading section of its own (illustrated by the red rectangle in the last column). The sequence-based logic is found to be fragile and does not resolve a fundamental property of the Schematismus' data logic. As illustrated in the second column of Figure 34, the Layout Model does not pick up the differences between headings at this level either.

The second approach was conceptualized by the Schematismus Project team in the form of a Neo4j graph, incorporating the predicted hierarchies. The examples demonstrate that this approach is unsuccessful due to the lack of granularity of the hierarchy levels of headings. As illustrated in the second column of Figure 34, the uniformity of the information hinders the identification of the correct linkages. The data obtained from the 'Curly' container element in example presented in chapter 5.3.3 demonstrated that knowledge about the end of a heading's zone of influence allows for the accurate modeling of the respective sections.

K. K. Hof-Musik-Capelle.	TOC-value	5
Hof-Capellmeister.	H5	1
	Paragraph	0
Vice-Hof-Capellmeister.	H5	1
	Paragraph	0
Titular-Vice-Hof-Capellmeister.	H5	1
	Paragraph	0
Hof- Sänger.	H5	2
Discantisten und Altisten.	H5	1
	Paragraph	0
Tenoristen.	H5	1
	Paragraph	0
Bassisten.	H5	1
	Paragraph	0
Hof-Musiker.	H5	2
Organisten.	H5	1
	Paragraph	0
Violinisten.	H5	1
...	Paragraph	0
Trompeter.	H5	1
	Paragraph	0
Pauker.	H5	1
	Paragraph	0
Hofmusik-Archivar.	H5	1
	Paragraph	0
Hofball-Musik-Director.	H5	1
	Paragraph	0

Figure 34: The table illustrates the insufficiency of information to correctly reconstruct the hierarchical linkages for the two last headings.

The fundamental problem lies in the fact that each heading possesses a certain zone of influence, yet the supplied data inadequately provides the necessary information concerning its extent. As recognizing the extent of those zones depends entirely on the correct identification of hierarchical differences between consecutive headings, the present master's thesis will continue with an evaluation of possibilities to improve the level of detail, with which headings are classified.

6 Assessment of Possible Improvements and Further Extractions

The apparent information deficit gives rise to the question regarding the means by which it might be reduced. It appears that a conclusive solution in the form of a complete reconstruction of the printed institutional structure can only be reached if the variation of heading styles can be deciphered completely. This can entail deconstructing the entire hierarchical code of heading styles of the Schematismus. However, it would likely be sufficient to determine for every heading if the preceding or following headings employ the same typography.

6.1 Evaluating the Reliability of Different Information Sources

The layout elements that are currently extracted by the layout model are inadequate for the conclusive reconstruction of the institutional structure of the Schematismus. Further layout-based extractions from the print could potentially yield additional information. The print layout contains a number of typographic elements that have not yet been thoroughly examined and which have the potential to yield valuable additional information. A preliminary evaluation of these elements will ascertain their reliability and information content.

A. Landes-Vertretung
B. Landes-Verwaltung
- I. Politische Behörden
  - K. K. Statthalterei/Landesregierung
    - Politische Bezirks-Behörden
    - Gemeinde-Behörden
- II. Unterrichts-Behörden und Anstalten
  - Landes-Schul-Rath/-Behörden
- III. Justiz-Behörden
- IV. Finanz-Behörden
  - Finanz-Direction
- V. Behörden für Handel und Verkehr
- VI. Behörden für Landescultur und Bergwesen
- VII. Behörden für Landesvertheidigung
  - Landes-Gendarmerie-Commando

C. Kirchliche Behörden
D. Öffentliche Institute

Figure 35: Recurring headings within the administration of these Kingdoms and Lands Represented in the Imperial Council

For this task, a repetitive section of the Schematismus was chosen. In the 1878 edition, the Cisleithanian half of the monarchy was divided into 14 administrative regions. The local administration of these “Kingdoms and Lands Represented in the Imperial Council” had a

consistent institutional framework. This was reflected in the Schematismus through a recurring template of nested headings. Figure 35 lists the higher-level headings that appear in each of the 14 sections. The four headings with capital letter numbering represent the highest hierarchical layer. The institution “B. Landes-Verwaltung” has a first sublayer that is consistently numbered with Roman numerals. The other headings are not numbered. Their hierarchical level and nesting are indicated by their indentation and inferred from their respective heading styles.

A comprehensive evaluation of all 14 administrative units was conducted to ascertain recurring layout features that could facilitate the identification of changes in the hierarchy. A segment of the resulting Excel table is presented in Figure 36. The complete evaluation is appended to this thesis in Annex 2: Assessment Table. The table provides a comparative analysis of four distinct information sources. The four sources in question are as follows: the style of the heading, the printed separation lines, the hierarchical information from the table of contents, and the numbering of headings.

6.1.1 Heading Styles

The left half of the Assessment Table (Annex 2) is used to support the evaluation of the heading styles. Distinctive indentations serve to denote the intra-sectional diversity of heading styles. The magnitude of indentation is determined by the hierarchical structure

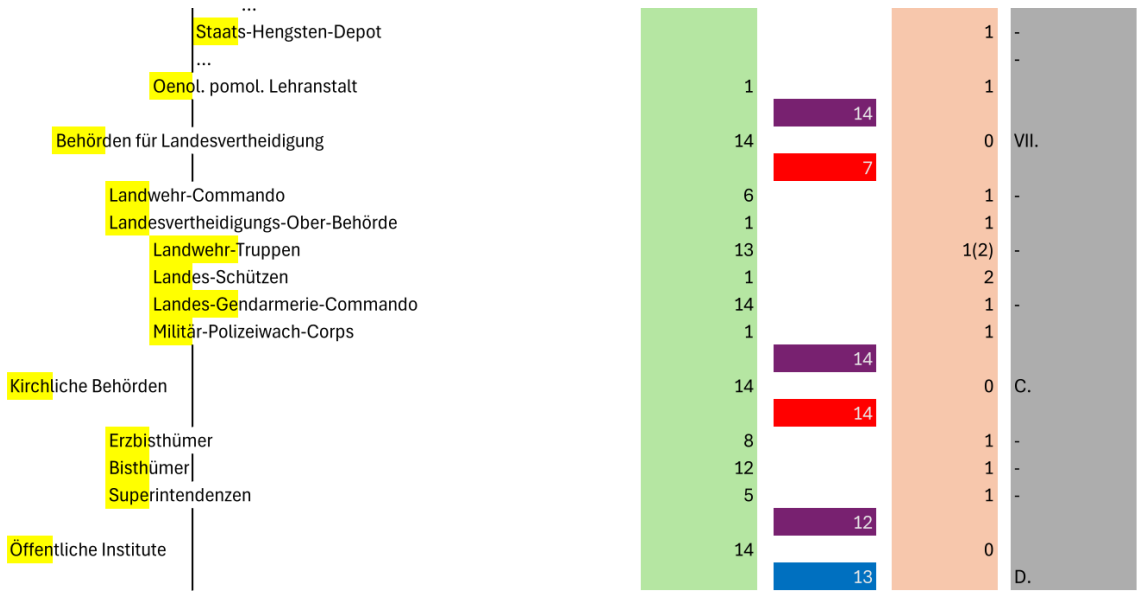


Figure 36: Screenshot from the evaluation table. The left part evaluates consistency and information content of heading styles, the green column counts the frequency of headings, the striped column evaluates separation lines, the orange column evaluates hierarchies in the table of contents, the grey column evaluates numberings of headings.

that emerges from the nested arrangement of elements. The color yellow is used as an indication of inter-sectional uniformity. This enables the visualization of whether a particular heading maintains a uniform style throughout all fourteen sections.

As illustrated in Figure 36, the heading style employed for “Landwehr-Truppen” and “Landes-Gendarmarie-Commando”, for instance, is not uniform, with two distinct styles being observed for each heading. The vertical line serves to denote the distinction between page-wide styles and the conventional three-column layout. Entries starting to the left of this delineation signify page-wide headings, while those to the right pertain to headings within columns. In instances where the yellow mark extends across the line the headings manifest in a page-wide configuration in certain sections and in a columnar arrangement in others.

The values in the green column are indicative of the frequency with which the respective heading occurs in the given section. A comprehensive transcription and enumeration of all page-wide headings, as well as the most frequently recurring higher-level headings in the columns, was conducted.

The evaluation indicates that the styles of the headings are the most significant and consistent informational units in the comparison. It is possible that slight discrepancies may be observed in the lower ranking hierarchical levels (i.e., within columns); however, within a single section, there is a high degree of consistency. Ascertaining the various styles of headings is a challenging endeavor, given the extensive range of styles that must be considered. However, it is the sole information source that can ensure precise results.

6.1.2 Separation Lines

Figure 37: Different types of separation lines and the color codes used in the evaluation table.

A notable aspect of the Schematismus layout has thus far gone unnoticed. Horizontal separation lines are a recurring feature in the main part of the later editions. Upon initial observation, it seems evident that the primary objective of the lines is to serve a structural role. They are intended to demarcate different sections of the Schematismus or indicate hierarchical break lines, where the heading alone is insufficiently distinct.

The occurrence of printed separation lines is evaluated in the third column of the Assessment Table (Annex 2). Within the sections under consideration, various styles of separation lines were detected. The various line styles are illustrated in Figure 37. The types differ in terms of length and ornamentation. The colored frames indicate the color code that was utilized to facilitate evaluation.

These lines manifest in various forms within the page-wide layout blocks and within columns. The present evaluation is constrained to the first five line variants, which exclusively appear in the page-wide layout. Within the three-column sections only the short line devoid of a color code is employed. The main area of usage of these shorter lines does not correspond with those repetitive sections that are evaluated through the Assessment Table. Therefore, the quantitative evaluation of this separation line variant is recommended to be conducted in future evaluations, should the present investigation find that separation lines in general could be a reliable source of structural information.

Readjusting the focus to the third column in the Assessment Table, an extract of which is shown in Figure 36, the line colors serve as visual cues, indicating which separation line variant was used to separate the respective headings of the left-side register. The number inscribed in the color coded cell serves as a quantitative metric, indicating the frequency with which the specific line variant is utilized. A value less than 14 indicates that the respective line variant was not employed in all 14 repetitive administrative sections under scrutiny. The instances that do not correspond to the rule may have employed alternative line variants or none at all.

As illustrated in the Assessment Table (Annex 2) and also reflected in Figure 36, the separation lines were employed with a high degree of consistency. Especially the separation lines employed for the higher-ranking institutions were consistent in their application and location. The exceptions from the rule demonstrate that the typesetting of the 19th century was dependent on human decisions and therefore occasionally flawed. For instance, the separation line preceding the element “B. Landes-Verwaltung” consistently exhibits the ornamented, magent-coded variant. Yet, in the province of Lower Austria, a simple straight line is employed, thus departing from the norm. In a similar fashion, the line preceding “Öffentliche Institute” in Galicia and Styria, and the line following the same section is Lower Austria depart from their respective norms.

The additional information value of the separation lines of page-wide layout blocks is limited, but relatively consistent throughout the sections under evaluation. The information could be useful in the differentiation of the highest hierarchical layers.

6.1.3 Table of Contents

The fourth column of the Assessment Table (Annex 2) collates data pertaining to hierarchical levels that can be deduced from the table of contents. The table of contents is first and foremost a reference table to navigate the printed content. However, by use of indentet lines its layout seemingly represents the hierarchy of the administrative units as well.

To find out, whether these indentations correspond with the perceived hierarchy of heading styles, the numerical values in the fourth column of the Assessment Table represent the indentations utilized in the table of contents for the corresponding headings to the left. In this system, the value of "0" is assigned to indicate a left-aligned entry, "1" is assigned to indicate a singular indent, and so forth. Should bracketed numbers be present, it is to be noted that the value within the brackets is employed less frequently than the value preceding the brackets.

As the short extract in Figure 36 is able to illustrate, the table of contents is unable to replicate the hierarchical variation represented by the headings and their nested structure in the main body of the Schematismus. The multitude of heading styles introduces a much greater level of detail of hierarchical levels than the three different indentation values of the table of contents could possibly represent. Using the table of contents as a source of reference, as was the case in the reconstruction approaches of the previous chapter, will result in erroneous hierarchies. In relation to the administrative institutions of the Kingdoms and Lands Represented in the Imperial Council, the table of contents establishes a uniform level for all headings of the first and second hierarchical layers. Consequently, the "Landes-Verwaltung" cannot be distinguished from its corresponding highest offices. In the event that a table of reference is required for the purpose of resorting a hierarchy, it is recommended that the necessary effort be invested to reconstruct it from the main part directly.

6.1.4 Numberings of Headings

The final column of the Assessment Table (Annex 2) is concerned with the evaluation of heading numberings. It has previously been observed that the two highest hierarchical layers feature headings that exhibit a consistent numbering sequence. The numbering styles and values demonstrate stability throughout the sections. It is evident that the

remaining headings under consideration do not employ numerical numbering systems. It is noteworthy that numberings are a common feature of lower-ranking headings, as well. In this case, it should be assumed that consistency extends only to the respective section. This is a qualitative assessment, as no rule-based inquiry has been conducted into those smaller headings. The numbering of headings offers a high degree of stability, but it must be noted that this also results in a limited added value. Nevertheless, they could be used for validation purposes.

Summing up the findings of the investigation conducted in this chapter it can be concluded that the hierarchical structure of the table of contents should not be used as a reference for actual organizational hierarchies of the administration. Heading numberings and separation lines can potentially add information to the extracts that are produced by the currently employed Layout Model, however the information gain is expected to be marginal and the means of extraction would need to be evaluated for both cases. Ultimately, only headings and their various styles yield the most reliable and conclusive information regarding hierarchical relationships between the administrative units contained in the Schematismus.

As has been argued in chapter 5, the data output currently produced by the layout model cannot conclusively reproduce the hierarchical information of the printed Schematismus. This is because the abstracted annotation schema is reduced to five heading categories and thus much simpler than the complexity of the print. However, it does not seem practical or even feasible to close the information gap utilizing a much finer grained annotation schema. Strategies to improve the level of detail with which different heading styles could be separated can relate to the introduction of post-processing steps performed on the structured output of the Layout Model or to the extraction of additional layout information.

6.2 Suggestions for Post-Processing Steps

In light of the conclusions drawn in the preceding section, the focus should be on considerations to improve the level of detail with which different heading styles can be differentiated. The subsequent discussion will present two post-processing approaches with the objective of addressing the information deficit.

6.2.1 *Section-Wise Clustering of Heading Styles*

It is acknowledged that page-wide sections contain headings of a higher order. It is therefore evident that the columns between two page-wide sections should coherently be subordinate to the preceding page-wide headings. This signifies that the columns and the preceding page-wide headings form a distinct branch of the organizational structure. It can be posited that, within the confines of this specific branch, the hierarchical values of different heading styles must be consistent.

The proposal is to utilize a clustering approach to the headings of such branches, based on visual characteristics. The output of the layout segmentation model can then be processed to extract the bounding box around every heading to be utilized as an image snippet. In an ideal scenario, these fragments could be categorized into as many distinct classes as there are unique heading styles in this particular section. The key to achieving this objective is to ascertain the optimum number of classes for each section. However, if the process of clustering the headings in distinct classes were to function as intended, this information could be utilized to facilitate the reconstruction of the organizational boundaries. The identification of whether consecutive headings or consecutive sections of headings and paragraphs are of the same or different class is of fundamental importance. This knowledge would allow us to make inferences regarding the limits of the zones of influence of headings.

6.2.2 *Perform Image Classification Tasks on Heading Snippets*

The process of section-wise clustering, which is based on visual characteristics, may encounter difficulties in determining the appropriate number of clusters. Applying image classification as a means of augmenting existing data with supplementary information could represent a more straightforward approach to discerning the headings of a given section. The additional information could be used to logically group the headings.

It seems probable that image classification algorithms are able to identify certain properties of the font. The Schematismus typography employs a combination of serif and non-serif fonts, along with bold and italic fonts, various font sizes and character spacings, to create a range of heading styles. One potential approach to address this issue would be to attempt to train an image classifier to disambiguate some of those properties. For instance, it appears feasible for a CNN model to accurately differentiate between serif and non-serif fonts. In such a scenario, the spatial information within the structured data output of the Layout Model would be used to extract image snippets for each heading individually. Those

snippets would then be processed by a trained image classifier to determine whether the depicted heading employs a serif font. The information produced by the classifier would then be added to the existing structured data and increase the information available.

6.3 Suggestions for Further Extraction Tasks

The enhancement of the information basis can be achieved either through post-processing existing data, or through the extraction of more data, or data of a higher quality. It is recommended that three distinct options for new extraction tasks be considered.

6.3.1 Separation Lines

There are at least five different types of separation lines. The majority of these are employed in the page-wide sections. Within columns, short straight lines appear sporadically to indicate a break within the same section.

To illustrate the function of this type of short horizontal line the short example concerning the “Hof-Musik-Capelle” that was discussed in chapter 5.3.2 (Figure 28) will be readopted. Regularly the significance of the listed positions declines from the beginning to the end of an institutional section. The initial listing is typically designated as a “Director” or a comparable position. In the case of the “Hof-Musik-Capelle”, the hierarchical structure comprises singers, musicians and other personnel, concluding with the Court Music Archivist. While this assertion may be met with dismay by scholars of history, it can be posited that the archivist’s role was the most subaltern position within the orchestra. Nevertheless, when evaluated in accordance with the structural logic of the sequential positioning of headings, the “Hofball-Musik Director” would be considered subordinate even to him. The person holding this title was Eduard Strauß, son of Johann Strauss I and member of the renowned Strauss dynasty. It is not possible to argue, on reasonable grounds, that this person should be ranked below the archivist. The only indication that the cascade of importance must be broken before the Musical Director is the short horizontal line.

Despite the absence of a definitive protocol for addressing short separation lines, the insights afforded by visual detection may prove to be of significant value in future endeavors. As demonstrated in the preceding section, the separation lines of the page-wide layout are employed with a high degree of consistency to differentiate the highest hierarchical layers. The information is consistent, and it is to be expected that the layout model will be able to recognize the lines with a high degree of accuracy.

6.3.2 *Improve Prediction Quality by Reducing Heading Classes*

It is evident that separation lines alone are insufficient for successful hierarchical reconstruction. Ultimately, the key lies in a fine-grained separation of heading types. The current annotation schema for the headings does not represent the full spectrum of stylistic variations and the existing predictions do not permit a full reconstruction of the organizational hierarchy. The training data was manually labelled by humans. Each page contains numerous headings, and thus the correct labelling of each heading consistently throughout a large number of pages is a challenging task. It is therefore very likely that the labeled training data includes some degree of inconsistency. Inconsistently labeled training data may lead to erroneous predictions, which in fact do occur despite the generally high degree of reliability of the Layout Model.

Therefore, it could be considered to reduce the complexity of the annotation schema. The existing headings could be relatively simply mapped automatically to a new annotation schema. It seems that the implementation of two heading classes could be efficacious. It is proposed that 'H0' should be employed for headings in page-wide layouts, while 'H1' should be used for headings in columns. The implementation of such an annotation schema would serve to perpetuate the prevailing hierarchical logic, while at the same time facilitating its straightforward compilation and control with a high degree of confidence.

The uniformity of the data could potentially further increase the quality of the predictions and facilitate post-processing tasks aimed at reconstructing the full granularity of the style information.

6.3.3 *Extract Heading Styles Using Fine Grained Annotation Schema*

In contradiction to the earlier assessment that a conclusively fine-grained annotation schema is not conceivable, a proposition is made in the following to experiment with such an approach, nevertheless. Although not very likely, a successful outcome would solve the issues that complicate the extraction hierarchical relationships in a single step. It is clear that adding post-processing steps would be unnecessary if the layout segmentation could successfully generate heading classes with enough detail to cover all possible heading styles.

However, this approach would require a complete relabeling of the training data. The current Layout Model was trained on a set of over 1,000 pages of training data that was manually labeled. The model can differentiate between headings across five categories. The new training data must be large enough to ensure reliable separation of all heading

types. This requires a lot of pages, which must be chosen carefully to include each heading style. Making such a dataset by hand would require a lot of effort and probably wouldn't result in the necessary accuracy and uniformity. Additionally, the probability of the machine learning model reproducing such a fine-grained categorization is low.

It is therefore not recommended to use manually labeled training data to implement such a complex annotation schema. Given the capacity to reliably segment all headings, irrespective of their type and hierarchy, image snippets of individual headings can be easily produced. These snippets could then be clustered according to their distinct visual characteristics. This approach is similar to the previous clustering approach, but it doesn't rely as much on the exact number of clusters. Provided the number of clusters is sufficiently large, a highly detailed annotation of heading classes is obtained. Ultimately, it is not necessary to discriminate all different headings of the entire edition; rather, discrimination between headings is required only between page-wide layout blocks. Any additional heading style improves the likelihood that a closed section contains only different heading categories.

Allowing the distinctive criteria to be determined by a computer algorithm could prove beneficial in terms of consistency and reproducibility by the layout model. Automated approaches to relabel training data could be implemented with limited human effort, thus offering a convenient way to experiment and evaluate the potential of such an approach.

7 Conclusions

The Schematismus Project – a joint research endeavor of the University of Graz and Graz University of Technology, which is aimed at an automated reconstruction of the Schematismus as a digital database – forms the nucleus around which this thesis was built. Nevertheless, the present thesis is an independent piece of work. The connection to the Schematismus Project is primarily through the data input for the analytical evaluations of the later chapters. This data was produced by the project's custom trained Layout Model. In its analytical chapters the thesis focuses on very specific considerations regarding the reconstruction of hierarchical relationships between different entities of the Habsburg civil service. Despite this very specific focus – or maybe because of it – the present thesis had to embrace a variety of topics to offer contextual information. Departing from historical considerations about the "Hof- und Staatsschematismus" the study explored prosopography and data modeling in theory and by examples to subsequently return to the Schematismus with a focus on evaluating the results of the automated data extraction efforts. The concluding remarks will, therefore, start with a brief recapitulation of the main topics.

The thesis started with an introduction to the "Hof and Staatsschematismus" as a source for historical research. This introduction encompassed a brief overview of genesis and evolution of the Schematismus's contents and more importantly its visual appearance. The conclusion that the later editions of the series, especially the years between 1876 and 1918, exhibit no major adaptations to the visual characteristics, is picked up in later chapters to explain, why these particular editions and their layout are the primary focus of the initial project phase of the Schematismus Project and the present master's thesis.

In the immediately subsequent chapter prosopography, "the investigation of the common background characteristics of a group of actors in history",[127] was introduced both as a conceptual frame within which to orient the Schematismus Project and its anticipated output, and as methodological reference regarding the modeling of data for prosopographical databases. Furthermore, the chapter on prosopography was used to portray other prosopographical research projects in connection with the personnel of the Habsburg court and state and highlight potential synergies. Theoretical considerations and exemplary evidence

[127] STONE, Prosopography, 1971, p. 46.

from recent prosopographical projects led to the conclusion that graph databases offer many benefits for prosopographical data representation.

This realization was used to lead over to the first analytical section of this thesis. Chapter 4 started by explaining the technological basis of the Layout Model of the Schematismus Project as well as thoroughly analyzing the data output that it can produce in the form of distinct layout classes. The understanding of the expected data output – respectively the data input for any reconstruction effort – inspired the definition of a first preliminary data model for a future graph database implementation of the extracted information. The data model, which is based on the main properties of the data representation of the printed version, consists of three entity types and two relationship types. These different entities and relations can be combined into three standard relationships:

1. Person-Position Relationship (A named individual holds a specific administrative position):

$$\underset{Position}{< \text{Präsident} >} \quad \underset{Relation}{< \text{is occupied by} >} \quad \underset{Person}{< \text{Carl Härdtl} >}$$

2. Position-Institution Relationship (An administrative position exists within a specific institutional context):

$$\underset{Institution}{< \text{Ausschuss} >} \quad \underset{Relation}{< \text{has} >} \quad \underset{Position}{< \text{Präsident} >}$$

3. The Institution-Institution Relationship (An administrative unit exists within a larger institutional hierarchy):

$$\underset{Institution}{< \text{Niederösterreichische Advocaten} - \text{Kammer} >} \quad \underset{Relation}{< \text{has} >} \quad \underset{Institution}{< \text{Ausschuss} >}$$

Building on this understanding of how the data output of the Layout Model is structured and how the data should be prepared for a envisioned graph database, the remainder of the thesis aimed at determining whether the extracted information is sufficient to digitally reconstruct the hierarchical data representation of the printed version in detail. Deconstructing the logic of the sequence of headings and paragraphs within the extracted dataset, an experimental, rule-based approach was developed to translate this logic into hierarchical relationships. Since the information on the sequence alone does not discriminate between different types of headings, limited easy to retrieve context information in the

form of a manually curated copy of the table of contents was additionally introduced. The effort was compared to a similar approach by the team of the Schematismus Project. It was concluded that neither of the two approaches provides an acceptably accurate reconstruction of the organizational structure across the entire printed source. However, when looking at secluded branches of the administration, the resulting reconstructions are more promising.

In the subsequent chapter 6, possibilities to expand the information basis are explored. The considerations are backed by a thorough evaluation of different sources for hierarchical and structural information. Fourteen repetitive sections of the Schematismus of 1878 to assess the consistency and reliability of different layout elements and the table of contents regarding representation of hierarchy were compared. The evaluation's results can be summarized as follows:

- The table of contents should not be used as a reference for hierarchy. The hierarchical implications of line indentations are inconsistent and do not conform to the hierarchical structure of the main body of the Schematismus.
- Separation lines can provide relatively consistent boundaries between the highest hierarchical layers. Small separation lines within columns carry hierarchical meaning but the exact significance of these small lines has yet to be explored in detail. Separation lines are, however, an easy to retrieve and potentially valuable source of information.
- The fine-grained disambiguation of heading styles is the only reliable option to achieve a detailed reconstruction of the printed hierarchy relations.

The initial research question of this thesis is addressed based on the findings of chapters 5 and 6. The general expectation that large parts of the organizational structure can be reconstructed based on the Schematismus's layout is concurred but with some notable limitations. Indeed, the various nested headings almost exclusively carry the critical information regarding hierarchical relationships between administrative units that are represented through the layout. There are very scarce options for reference and validation. The printed separation lines serve this purpose only in a very limited number of instances and only with regard to determining institutional boundaries. The table of contents is not

reliable in terms of hierarchical structuring. It is emphasized that large parts of the organizational information could potentially be extracted from the layout but not the entire respective information. The considerations in chapter 5.2.4 serve to illustrate, that the 'Position' entity within the introduced data model is in many instances concealed within the 'Paragraph' layout class and can only be accessed via a semantic analysis.

Regarding the extent to which the hierarchical organizational structures can be reproduced from the concrete dataset of extracted layout information, the assessment is ambivalent. However promising the result on single pages and secluded branches may look, a conclusive reproduction of the structures of the entire printed work is not feasible. The three comparative examples in chapter 5.3 illustrate that both reconstruction efforts that were evaluated produce incorrect linkages on several occasions. The existing data lacks information on the extent of hierarchical structures. The beginning of sections can mostly be determined but not always their extent. As evidenced by the analysis in chapter 6.1 the only feasible strategy to solve the reconstruction task without accessing textual semantics, is a very detailed disambiguation of different heading styles.

Although the extracted information is not yet rich enough to reconstruct the hierarchy of the Schematismus consistently in full detail, there is reason for confidence that established technological approaches could potentially reduce the information deficit decisively. The suggestions made in chapters 6.2 and 6.3 for further extraction tasks highlight different strategies to achieve a more fine-grained disambiguation of headings. The strategies can either emphasize post-processing or aim at a more detailed annotation schema.

The presented post-processing approaches will be summarized in the following to give a promising outlook on further research. The approaches rely foremost on a reliable separation of headings and paragraphs. The performance of the Layout Model in this regard is very good. It would be worth evaluating if the performance could be increased even further by retraining on reduced heading classes as suggested in section 6.3.2. The snippets of individual headings, that can be produced using the extracted information, would then undergo post-processing. Two options for post-processing operations, image clustering and image classification, have been presented. When embracing the classification approach, several classifiers could be trained to discriminate serif from non-serif fonts, bold from regular and italic fonts, or direct categorization of the most frequent headings. The predicted classes could expand the existing information for each heading. Image classification

is a well-established task in computer vision and the approach would very likely yield additional parameters on which headings could be disambiguated.

8 List of References

Cited Publications

AKOKA Jacky / COMYN-WATTIAU Isabelle / LAMASSÉ Stéphane et al., Contribution of Conceptual Modeling to Enhancing Historians' Intuition -Application to Prosopography. In: International Conference on Conceptual Modeling (ER) Vienna, Austria, 2020.

ANDERMANN Kurt, Pragmatische Schriftlichkeit. In: Handbuch Höfe und Residenzen im spätmittelalterlichen Reich - Band15.III (= Residenzenforschung 15,3). Ostfildern: Thorbecke, 2007, pp. 37–60.

ANDREWS Tara, The STructured Assertion Record (STAR) Model for Event-based Representation of Historical Information. Mainz 2023. Online reference: graphentechnologien.hypotheses.org/files/2023/05/GrapHNR-2023-32-Andrews-STAR.pdf, (last accessed: 08.02.2025).

ANDREWS Tara / RÓZSA Márton / PRAJDA Katalin et al., Re-Evaluating the Eleventh Century through Linked Events and Entities. In: Historical Studies on Central Europe 4, 1, 2024, pp. 217–245.

BAILLIE James, Alternative Database Structures for Prosopographical Research. In: International Journal of Humanities and Arts Computing 15, 1-2, 2021, pp. 117–132.

BAUER Volker, Territoriale Amtskalender und Amtshandbücher im Alten Reich. In: Rechtsgeschichte - Legal History 2002, 01, 2002, pp. 71–89.

BERNÁD Ágoston Zénó / LEJTOVICZ Katalin, Korpusanalyse und digitale Quellenkritik – Die Vermessung des Österreichischen Biographischen Lexikons. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS). Vom gedruckten Textkorpus zur Webapplikation für die Forschung. Wien – Hamburg: new academic press, 2020, pp. 107–158.

BOILLET Melodie / KERMORVANT Christopher et al., Multiple Document Datasets Pretraining Improves Text Line Detection With Deep Neural Networks 2020. Online reference: arxiv.org/pdf/2012.14163, (last accessed: 08.02.2025).

BOILLET Mélodie / TARRIDE Solène et al., The Socface Project: Large-Scale Collection, Processing, and Analysis of a Century of French Censuses 2024. Online reference: arxiv.org/pdf/2404.18706v2, (last accessed: 08.02.2025).

BRADLEY J., Texts into Databases: The Evolving Field of New-style Prosopography. In: Literary and Linguistic Computing 20, Suppl 1, 2005, pp. 3–24.

BRANDT Harm-Hinrich, Der österreichische Neoabsolutismus. Staatsfinanzen und Politik; 1848 - 1860. Zugl.: München, Univ., 09 - Fachbereich Geschichts- u. Kunstwiss., Habil.-Schr., 1974/75 (=Schriftenreihe der Historischen Kommission bei der Bayerischen Akademie der Wissenschaften 15). Göttingen: Vandenhoeck und Ruprecht 1878.

CARTER Philip, What is National Biography For? Dictionaries and Digital History. In: Karen FOX (Ed.) True Biographies of Nations? The Cultural Journeys of Dictionaries of National Biography: ANU Press, 2019, pp. 57–78.

CONSTUM Thomas / TRANOUEZ Pierrick et al., DANIEL: A fast Document Attention Network for Information Extraction and Labelling of handwritten documents 2024. Online reference: arxiv.org/pdf/2407.09103v1, (last accessed: 08.02.2025).

COQUENET Denis / CHATELAIN Clement / PAQUET Thierry, DAN: A Segmentation-Free Document Attention Network for Handwritten Document Recognition. In: IEEE transactions on pattern analysis and machine intelligence 45, 7, 2023, pp. 8227–8243.

CROFTS Nicholas / DOERR Martin et al., The CIDOC Conceptual Reference Model: A standard for communicating cultural contents. Online reference: cidoc.ics.forth.gr/docs/martin_a_2003_comm_cul_cont.htm, (last accessed: 01.02.2025).

CYGANIAK Richard / WOOD David et al., RDF 1.1 Concepts and Abstract Syntax 2014. Online reference: www.w3.org/TR/2014/REC-rdf11-concepts-20140225/, (last accessed: 01.02.2025).

DEAK John, Forging a multinational state. State making in imperial Austria from the Enlightenment to the First World War (=Stanford studies on Central and Eastern Europe). Stanford, California: Stanford University Press 2015.

DIEM Markus / KLEBER Florian / FIEL Stefan et al., cBAD: ICDAR2017 Competition on Baseline Detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): IEEE, 2017, pp. 1355–1360.

FEIGL Roland / GRUBER Christine, Das Projekt APIS und der ÖBL-Textkorpus in seiner digitalen Transformation. Herausforderungen für ein traditionelles biographisches Lexikon. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS). Vom gedruckten Textkorpus zur Webapplikation für die Forschung. Wien – Hamburg: new academic press, 2020, pp. 9–18.

FLEISCHHACKER David, Object Detection for Improved OCR via Faster R-CNNs on Historical Schematism-State Manuals. Bachelor's Thesis, Graz University of Technology. Graz.

FLEISCHHACKER David / GOEDERLE Wolfgang et al., Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach 2024. Online reference: arxiv.org/pdf/2401.07787, (last accessed: 08.02.2025).

GÖDERLE Wolfgang, Workshop - Extracting the Hof- und Staatshandbuch. Graz 2024.

HAMDAN Mohammed / RAHICHE Abderrahmane et al., HAND: Hierarchical Attention Network for Multi-Scale Handwritten Document Recognition and Layout Analysis 2024. Online reference: arxiv.org/pdf/2412.18981v1, (last accessed: 08.02.2025).

HEINDL Waltraud, Gehorsame Rebellen. Bürokratie und Beamte in Österreich 1780 bis 1848 (=Studien zu Politik und Verwaltung 36). Wien: Böhlau 1990.

HOTSON Howard / WALLNIG Thomas / JENSEN Mikkel Munthe et al., Prosopographies of the Republic of Letters. In: Howard HOTSON / Thomas WALLNIG (Eds.) Reassembling the republic of letters in the digital age. Standards, Systems, Scholarship. Göttingen: Universitätsverlag Göttingen, 2019, pp. 371–398.

HUSSAIN Mahbub / BIRD Jordan J. / FARIA Diego R., A Study on CNN Transfer Learning for Image Classification. In: Ahmad LOTFI (Ed.) Advances in Computational Intelligence Systems. Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK (= Advances in Intelligent Systems and Computing Ser v.840). Cham: Springer, 2019, pp. 191–202.

JANNIDIS Fotis, Grundlagen der Datenmodellierung. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 99–108.

JANNIDIS Fotis, Netzwerke. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 147–161.

KAISER Maximilian / ROMBERG Marion / SCHLÖGL Matthias et al., Von APIS zu VieCPro. In: Helmuth ALBRECHT / Michael FARRENKOPF / Helmut MAIER et al. (Eds.) Historische Biographik und kritische Prosopographie als Instrumente in den Geschichtswissenschaften (= Veröffentlichungen Aus Dem Deutschen Bergbau-Museum Bochum 257). Berlin – Boston: De Gruyter Oldenbourg, 2023, pp. 117–140.

KANG Le / KUMAR Jayant / YE Peng et al., Convolutional Neural Networks for Document Image Classification. In: 2014 22nd International Conference on Pattern Recognition: IEEE, 2014, pp. 3168–3172.

KEATS-ROHAN Katharine S. B., Biography, Identity and Names: Understanding the Pursuit of the Individual in Prosopography. In: Katharine S. B. KEATS-ROHAN (Ed.) Prosopography approaches and applications. A handbook (= Prosopographica et Genealogica 13). Oxford: Unit for Prosopographical Reasearch Linacre College Univ. of Oxford, 2007, pp. 139–182.

KLINKE Harald, Datenbanken. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 109–127.

KRIZHEVSKY Alex / SUTSKEVER Ilya / HINTON Geoffrey E., ImageNet classification with deep convolutional neural networks. In: Communications of the ACM 60, 6, 2017, pp. 84–90.

KUBISKA-SCHARL Irene / PÖLZL Michael, Adelige und bürgerliche Karrierewege bei Hof. Eine Prosopographie des Wiener Hofpersonals 1711–1806. In: Akademie der Wissenschaft zu Göttingen (Ed.) Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen. Neue Folge: Stadt und Hof; 3: Göttingen Academy of Sciences, 2014, pp. 76–86.

KURZ Stephan, oeaw-ministerratsprotokolle/mp-edition-data: v. 1.5 including CMR calendar data 1872–1914. Wien 2024. Online reference: zenodo.org/records/11484662, (last accessed: 08.02.2025).

KURZ Stephan / FISCHER-NEBMAIER Wladimir / KAMPKASPAR Dario et al., Die Edition der Ministerratsprotokolle 1848-1918 digital. Workflows, Möglichkeiten, Grenzen. In: digital humanities austria 2018. empowering researchers. Vienna: Austrian Academy of Sciences Press, pp. 83–90.

LEJTOVICZ Katalin / SCHLÖGL Matthias, Die (semi-)automatische Verarbeitung biographischer Artikel im APIS-Framework. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS). Vom gedruckten Textkorpus zur Webapplikation für die Forschung. Wien – Hamburg: new academic press, 2020, pp. 49–82.

Neo4j Graph Database - Product Brief. The Most Trusted Database for Intelligent Applications 2022. Online reference: go.neo4j.com/rs/710-RRC-335/images/Neo4j-product-brief-database-US-EN.pdf, (last accessed: 02.02.2025).

NETWORKX DEVELOPERS, NetworkX 2024. Online reference: networkx.org/, (last accessed: 02.02.2025).

NOFLATSCHER Heinz, Hofstaatsverzeichnisse und Hof- und Staatsschematismen. In: Handbuch Höfe und Residenzen im spätmittelalterlichen Reich - Band15.III (= Residenzenforschung 15,3). Ostfildern: Thorbecke, 2007, pp. 409–432.

NOFLATSCHER Heinz, Ordonnances de l'hôtel, Hofstaatsverzeichnisse, Hof- und Staatskalender. In: Josef PAUSER / Martin SCHEUTZ / Thomas WINKELBAUER (Eds.) Quellenkunde der Habsburgermonarchie. (16. - 18. Jahrhundert) ; ein exemplarisches Handbuch (= Mitteilungen des Instituts für Österreichische Geschichtsforschung Ergänzungsband 44). Wien – München: Oldenbourg, 2004, pp. 59–75.

OLDMAN Dominic / DOERR Martin / GRADMANN Stefan, Zen and the Art of Linked Data. In: Susan SCHREIBMAN / Raymond George SIEMENS / John UNSWORTH (Eds.) A new companion to digital humanities (= Blackwell companions to literature and culture 93). Chichester, West Sussex – Malden, MA et al.: Wiley Blackwell, 2016, pp. 251–273.

OSTERHAMMEL Jürgen, Die Verwandlung der Welt. Eine Geschichte des 19. Jahrhunderts (=Historische Bibliothek der Gerda Henkel Stiftung). München: C.H. Beck [5] 2010.

PASIN Michele / BRADLEY John, Factoid-based prosopography and computer ontologies: towards an integrated approach. In: Digital Scholarship in the Humanities 30, 1, 2015, pp. 86–97.

REHBEIN Malte, Digitalisierung. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 179–198.

REHBEIN Malte, Ontologien. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 162–176.

ROMBERG Marion / KAISER Maximilian, The Viennese Court. A Prosopographical Portal. Eine Projektvorschau auf ein Referenz- und Nachschlageportal zum Wiener Hof von Leopold I. bis Franz II. (I.). In: Akademie der Wissenschaft zu Göttingen (Ed.) Mitteilungen der Residenzen-Kommission der Akademie der Wissenschaften zu Göttingen. Neue Folge: Stadt und Hof; 9: Göttingen Academy of Sciences, 2020, pp. 43–52.

RUMPLER Helmut, Integration und Modernisierung. Der historische Ort des „Neoabsolutismus" in der Geschichte der Habsburgermonarchie. In: Harm-Hinrich BRANDT (Ed.) Der österreichische Neoabsolutismus als Verfassungs- und Verwaltungsproblem.

Diskussionen über einen strittigen Epochenbegriff (= Veröffentlichungen der Kommission für Neuere Geschichte Österreichs Band 108). Wien – Köln et al.: Böhlau, 2014, pp. 73–81.

SCHEWE Gerhard, Organigramm 2018. Online reference: wirtschaftslexikon.gabler.de/definition/organigramm-44460/version-267771, (last accessed: 01.02.2025).

SCHLÖGL Matthias, APIS ER data model 2019. Online reference: zenodo.org/records/3568085, (last accessed: 08.02.2025).

SCHLÖGL Matthias / LEJTOVICZ Katalin, Die APIS-(Web-)Applikation, das Datenmodell und System. In: Christine GRUBER / Josef KOHLBACHER / Eveline WANDL-VOGT (Eds.) The Austrian Prosopographical Information System (APIS). Vom gedruckten Textkorpus zur Webapplikation für die Forschung. Wien – Hamburg: new academic press, 2020, pp. 31–48.

SCHÖCH Christof, Ein digitales Textformat für die Literaturwissenschaft: Die Richtlinien der Text Encoding Initiative und ihr Einsatz bei Textkonstitution und Textanalyse. In: Romanische Studien 1, 4, 2016, pp. 325–364.

SCHÖCH Christof, Information Retrieval. In: Fotis JANNIDIS / Hubertus KOHLE / Malte REHBEIN (Eds.) Digital Humanities. Eine Einführung. Stuttgart: J.B. Metzler Verlag, 2017, pp. 268–298.

SHEHZADI Tahira / STRICKER Didier et al., A Hybrid Approach for Document Layout Analysis in Document images 2024. Online reference: arxiv.org/pdf/2404.17888v2, (last accessed: 08.02.2025).

SHEIKH Talha Uddin / SHEHZADI Tahira et al., UnSupDLA: Towards Unsupervised Document Layout Analysis 2024. Online reference: arxiv.org/pdf/2406.06236v1, (last accessed: 08.02.2025).

SMITH R., An Overview of the Tesseract OCR Engine. In: Proceedings / Ninth International Conference on Document Analysis and Recognition. Curitiba, Paraná, Brazil, September 23 - 26, 2007. Los Alamitos, Calif.: IEEE Computer Soc, 2007, pp. 629–633.

STONE Lawrence, Prosopography. In: Daedalus 100, 1, 1971, pp. 46–79.

TARRIDE Solène / MAARAND Martin / BOILLET Mélodie et al., Large-scale genealogical information extraction from handwritten Quebec parish records. In: International Journal on Document Analysis and Recognition (IJDAR) 26, 3, 2023, pp. 255–272.

THE PANDAS DEVELOPMENT TEAM, pandas-dev/pandas: Pandas 2024. Online reference: pandas.pydata.org/, (last accessed: 02.02.2025).

VERBOVEN Koenraad / CARLIER Myriam / DUMOLYN Jan, A Short Manual to the Art of Prosopography. In: Katharine S. B. KEATS-ROHAN (Ed.) Prosopography approaches and applications. A handbook (= Prosopographica et Genealogica 13). Oxford: Unit for Prosopographical Reasearch Linacre College Univ. of Oxford, 2007, pp. 35–70.

WANG Chien-Yao / YEH I-Hau et al., YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information 2024. Online reference: arxiv.org/pdf/2402.13616v2, (last accessed: 08.02.2025).

WANG Jiawei / HU Kai et al., DLAFormer: An End-to-End Transformer For Document Layout Analysis 2024. Online reference: arxiv.org/pdf/2405.11757v1, (last accessed: 08.02.2025).

WIDDER Ellen, Hofordnungen. In: Handbuch Höfe und Residenzen im spätmittelalterlichen Reich - Band15.III (= Residenzenforschung 15,3). Ostfildern: Thorbecke, 2007, pp. 391–408.

ZHAO Zhiyuan / KANG Hengrui et al., DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception 2024. Online reference: arxiv.org/pdf/2410.12628v1, (last accessed: 08.02.2025).

ZONG Zhuofan / SONG Guanglu et al., DETRs with Collaborative Hybrid Assignments Training 2022. Online reference: arxiv.org/pdf/2211.12860v6, (last accessed: 08.02.2025).

Further References to Online Sources

ALEX: https://alex.onb.ac.at/static_tables/shb.htm (last accessed: 21.07.2025)

BOUNDING BOX EDITOR: https://github.com/mfl28/BoundingBoxEditor (last accessed: 02.02.2025)

CIDOC-CRM: https://cidoc-crm.org/ (last accessed: 02.05.2025)

GND: https://gnd.network/Webs/gnd/DE/Home/home_node.html (last accessed: 02.02.2025)

MPR: https://mpr.acdh.oeaw.ac.at/ (last accessed: 02.02.2025)

RELEVEN: https://releven.univie.ac.at/ (last accessed: 11.05.2025)

WIKISOURCE: https://de.wikisource.org/wiki/August_Friedrich_Pauly#Herausgeberschaft (last accessed: 01.02.2025)

Declaration on the Use of Artificial Intelligence

I hereby declare that artificial intelligence tools have been used in a limited capacity during the preparation of this master's thesis for the following specific purposes:

1. Translation assistance: AI tools, such as DeepL, ChatGPT or Claude AI were employed in a limited capacity to translate between German and English as well as to proof-read selected passages.
2. Code documentation: The AI tool "GitHub CoPilot" was used to generate explanatory comments for Python code included in the thesis appendix to enhance readability and understanding.

The core research, analysis, argumentation, and writing of this thesis remain my original work. The intellectual contribution, critical thinking, and scholarly conclusions presented in this work are entirely my own.

9 Annex 1: Commented Code Repository

Universität Wien

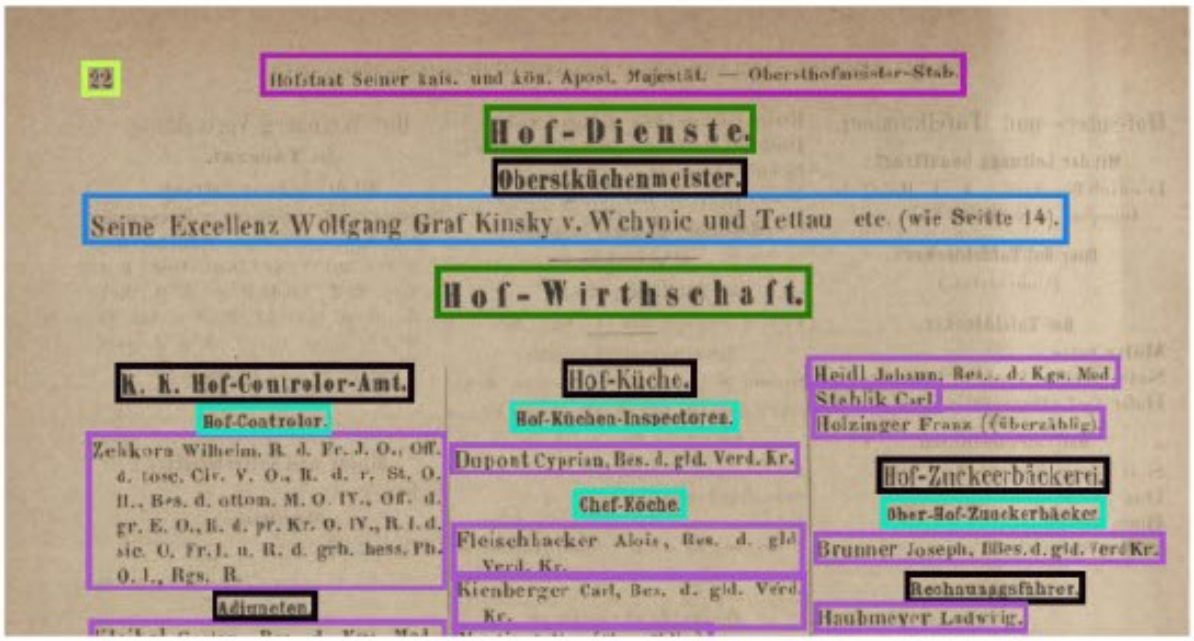
Master Thesis - Commented Code Repository

Reconstructing the Habsburg Office Structure from Data extracted from the 1878 Schematismus

The notebook tries to recover a hierarchical organizational tree-structure that is depicted in the Schematismus of 1878. "Schematismus" is the term commonly used for periodically printed almanachs of the civil service of the Habsburg Monarchy, published between 1702 and 1918. CSV files that contain information that is received from a YOLOv9-based layout detection model and a finetuned OCR engine serve as the input [1]. There is one CSV file per printed page. I try to find a rule-based heuristic to reconstruct the organizational tree-structure based on the information contained in the csv-data.

Ui	index	image_name	X_mir	Y_mir	X_ma	Y_ma	class	conf	pred_text
0	0	page-0196.jpg	0.185169	0.088071	0.206361	0.100887	Pagenumber	0.9070	18
1	1	page-0196.jpg	0.317659	0.087855	0.809880	0.101124	Header	0.9116	Hofstaat Seiner kais. und kön. Apost. Majestät. - Obersthofmeister-Stab.
2	2	page-0196.jpg	0.183339	0.112674	0.430726	0.137095	Paragraph	0.9761	Protiwensky Wilhelm, Bes. d. Kgs. Med. u. d. pers. S. u. L. O. V.
3	3	page-0196.jpg	0.183417	0.136516	0.430861	0.161242	Paragraph	0.9692	Kállay Béla v. Nagy-Kálló, Bes. d. Kgs. Med. (prov.).
4	4	page-0196.jpg	0.242330	0.164891	0.375470	0.176865	H5	0.9069	Hof-Fourier-Ansager.
5	5	page-0196.jpg	0.183098	0.178672	0.392373	0.190687	Paragraph	0.9681	Latty Jacob, Bes. d. Kgs. Med.
6	6	page-0196.jpg	0.245441	0.194747	0.372621	0.206597	H5	0.9199	Hof-Stabs-Adjutant.
7	7	page-0196.jpg	0.183080	0.208473	0.430961	0.220899	Paragraph	0.9662	Klinger Johann, Oberlt., wie Seite 27.
8	8	page-0196.jpg	0.238143	0.224528	0.376051	0.236520	H5	0.9004	Hof-Stabs-Feldwebel.
9	9	page-0196.jpg	0.183962	0.238222	0.431978	0.262305	Paragraph	0.9756	Troszt Anton, Bes. d. slb. T. M. I., d. Kgs. Med. u. d. r. G. Kr. IV.
10	10	page-0196.jpg	0.183846	0.261912	0.432321	0.298112	Paragraph	0.9761	Hauffe August, Bes. d Kgs. Med., d. r. gld. M. (a. St. Bd.) u. d. slb. M. d. pers. S. u. L. O.

The above image shows a snippet from the input data. Each line in the file holds information relating to a single layout element. As can be seen, the data is structured in nine columns. The first two are index-values, the third contains the name of the source image file, columns four through seven locate the layout element on the scanned page. Column eight gives the assigned class of the layout element, column nine a confidence value of the layout prediction and column ten holds the text contained in the layout snippet as predicted by the OCR-engine.



The different layout elements encode for different text blocks. The data in the CSV files contains classes for the page header and page number, six classes to separate different headings and two classes to distinguish running text. The heading classes H0, H1, H2, H4, H5 and H6 to some degree encode hierarchy, but one has to be cautious and cannot fully rely on the implications given by the class assignment. Generally, H0 is thought to have the highest Hierarchy and H5 the comparatively lowest. H6 is used for additional information given in brackets often belonging to the preceding heading. The list-

like entries between headings each identify a person holding a position in the Habsburg administration. Each such entry is labeled as "Paragraph", if within the common table structure of the Schematismus or as "BigParagraph" if the Person appears as a page-wide layout element.

Pandas is the library that I mainly use for the work on the data. All other imports serve special purposes. The networx library is used to produce the visualizations.

In [3]:
```
import pandas as pd
import matplotlib
import numpy as np
import os
import glob

import networkx as nx
```

I intended to start on a single page that is relatively simply structured. I wanted to understand the structural logic of the listings and try to find a way to solve the problem of hierarchical resorting. If a practical solution for the first page could be found, I would take my approach to the next page and evaluate there and evolve the code as necessary. This iterative approach is not easily depicted in a single notebook. I will rather show a unified version of my code that has been evolved and adapted over three cycles.

Preparing the Pandas Dataframe

With the first lines of code, I wanted to remove unnecessary and redundant information from the data. The pipeline was set up to process multiple consequent pages and developed on a 189-pages long sub-sample of the data set. The code iterates through the files in a given directory. On each page it looks for the page number information and evaluates the correctness of the respective "pred_text"-value. The numerical value is considered correct if it increses the value on the respectively preceding page by 1. After these checks the information is reduced, a new index set, and the data of each page is appended to a list data type.

In [12]:
```
path_of_the_directory = 'ocr'

to_drop = ["Unnamed: 0", "conf", "image_name", "X_min", "Y_min", "X_max", "Y_max"]
classes_to_keep = ["H0", "H1", "H2", "H3", "H4", "H5", "H6", "BigParagraph", "Paragraph"]
type_paragraph = ["BigParagraph", "Paragraph"]
type_heading = ["H0", "H1", "H2", "H3", "H4", "H5"]

# the page number of the preceding page is the start value for the page number evaluation
prev_page = 12

# the code reads all csv files in the directory evaluates the page number and concatenates
# the pages to a single dataframe
ocr_pages = []
for filename in sorted(glob.iglob(f'{path_of_the_directory }/*.csv')):
    df = pd.read_csv(filename)
```

```
    # evaluate if the page number is correct and then remove all data that is not needed
    # there should only one page number in the file and it should be the next page number
    # the index is set to a page-specific value
    page = prev_page
    if df[df["class"] == "Pagenumber"]["pred_text"].count() == 1:
        if int(df[df["class"] == "Pagenumber"]["pred_text"][0]) - page == 1:
            page = int(df[df["class"] == "Pagenumber"]["pred_text"][0])
            df = df.loc[df['class'].isin(classes_to_keep)]
            df = df.drop(to_drop, axis=1)
            #df = df.dropna(how="any")
            df["index"] = df["index"].astype(dtype="string")
            df["index"] = str(page)+"_"+df["index"]
            df["page"] = str(page)
        else:
    # allow manual correction of wrong page numbers by raising exceptions
            raise Exception(f"error in file {filename}. irregularity in page count. last corre
                            page: {page}")

    else:
        raise Exception(f"error in file {filename}. Multiple Page Numbers in file {filename}")

    ocr_pages.append(df)
```

The following are three error messages indicating wrong page numbers that occurred during the evaluation of 189 pages. Most frequently either the page number was not recognized as a number, or as a wrong number. When a message appeared, I went back to the respective CSV file and corrected the error.

The error message in the first image indicates that the count doesn't match expectation. An integer is correctly recognized but not the correct one. The error in the second image related to page 129. Obviously the "9" was not resolved correctly by the OCR. The very strange exception #3 appeared on pages 149, 167, 171. Although there is only one page number printed, the OCR recognizes double page numbers. There were nine error messages in the 189-page subsample.

```
Exception                                 Traceback (most recent call last)
Cell In[4], line 25
     23         df["page"] = str(page)
     24     else:
---> 25         raise Exception(f"irregularity in page count. last correct page: {page}")
     27 else:
     28     raise Exception(f"Multiple Page Numbers in file {filename}")

Exception: irregularity in page count. last correct page: 47

ValueError                                Traceback (most recent call last)
Cell In[4], line 16
     13 df = pd.read_csv(filename)
     15 if df[df["class"] == "Pagenumber"]["pred_text"].count() == 1:
---> 16     if int(df[df["class"] == "Pagenumber"]["pred_text"][0]) - page == 1:
     17         page = int(df[df["class"] == "Pagenumber"]["pred_text"][0])
     18         df = df.loc[df['class'].isin(classes_to_keep)]

ValueError: invalid literal for int() with base 10: '12)'

ValueError                                Traceback (most recent call last)
Cell In[3], line 16
     13 df = pd.read_csv(filename)
     15 if df[df["class"] == "Pagenumber"]["pred_text"].count() == 1:
---> 16     if int(df[df["class"] == "Pagenumber"]["pred_text"][0]) - page == 1:
     17         page = int(df[df["class"] == "Pagenumber"]["pred_text"][0])
     18         df = df.loc[df['class'].isin(classes_to_keep)]

ValueError: invalid literal for int() with base 10: '149\n149'
```

For the second step, I first concatenated the list elements to a single dataframe. The new dataframe was called "court_df" because the 189 pages of the sample belong to the chapter depicting the Habsburg Court.

The H6 annotation classes are not headings in the narrower sense. In fact, the H6 classes most commonly are neither institutions nor persons and thus hinder the reconstruction of hierarchy more than they help. Therefore, we want to keep the information but eliminate the H6 classes from the flow of elements belonging to the hierarchical structure.

The 147 appearances of H6 labeled Elements on the first 189-page subsample were therefore reclassified manually. This was done in a separate CSV file only containing the H6-elements. Introducing a new column, I assigned each H6 instance one of six categories. The manually corrected H6 data was imported into the concatenated dataframe and each category of H6 dealt with according to specifications.

In [13]:
```
# concatenate all pages to a single dataframe, reset the index and drop unnecessary columns
court_df = pd.concat(ocr_pages)
court_df = court_df.reset_index(drop=True)
court_df = court_df[["page", "index", "pred_text", "class"]]

# import the manual corrections for the class H6 and join them to the dataframe
class_h6 = pd.read_csv("H6_class_correctedMAN.csv", index_col=0).drop(["index",
                                                                       "class",
                                                                       "pred_text"], axis=1)
court_df = court_df.join(class_h6)


pd.options.mode.copy_on_write = True
# handle case pers_equiv: the class is set to Paragraph
court_df.loc[court_df['H6_type'] == "pers_equiv", "class"] = "Paragraph"

# handle cases address and info by creating new columns that contain the respective text
court_df.loc[court_df['H6_type'] == "address", "address"] = court_df.loc[court_df['H6_type'] =
```

```
                                                                    "address", "pred_text
court_df.loc[court_df['H6_type'] == "info", "info"] = court_df.loc[court_df['H6_type'] ==
                                                                    "info", "pred_text"]

# these new columns need to move a row up, as it is additional information
# to the preceding element
court_df["address"] = court_df["address"].tolist()[1:] + court_df["address"].tolist()[:1]
court_df["info"] = court_df["info"].tolist()[1:] + court_df["info"].tolist()[:1]

# drop the H6 rows that are not needed anymore
court_df = court_df.loc[court_df['H6_type'].isin(["address", "info", "none"]) == False]
```

The H6 category "mistake" was handled in a separate dataframe manually. This category was assigned once among the 147 H6 cases. Handling this single mistake by rule is technological overkill. It was performed as exercise. In these cases, the text wrongly classified as H6 belonged to the previous element and should be appended to the "pred_text" string of the previous row.

[14]:
```
court_df.loc[court_df.H6_type == "mistake"]
```

[14]:

	page	index	pred_text	class	H6_type	address	info
2228	36	36_63	versehen.)	H6	mistake	NaN	NaN

[ ]:
```
# handle the case mistake by concatenating the text of the mistake with the preceding text
# get the text of the mistake
search_text = court_df.loc[court_df.H6_type == "mistake", "pred_text"].values[0]
#define a new column that contains the text of the mistake
court_df.loc[court_df.H6_type == "mistake", "mistake"] = court_df.loc[court_df.H6_type ==
                                                                    "mistake", "pred_text"]
# move the mistake column one row up
court_df["mistake"] = court_df["mistake"].tolist()[1:] + court_df["mistake"].tolist()[:1]

# concatenate the text of the mistake with the preceding text
court_df.loc[court_df.mistake == search_text, "pred_text"] = str(
    court_df.loc[court_df.mistake == search_text, "pred_text"].values[0]) +" "+ str(
        court_df.loc[court_df.mistake == search_text, "mistake"].values[0])
# drop the mistake column
court_df = court_df.drop("mistake", axis=1)

# drop the rows that are not needed anymore
court_df = court_df.loc[court_df['H6_type'].isin(["mistake"]) == False]
```

The sixth H6 category named "categorization" actually has the function of a heading and therefore stays in the dataframe and also retains its class H6. With this, all H6 types of the first 189 pages have been taken care of. Each line of the dataframe we receive represents an element that is part of the hierarchical, organizational structure we are after.

Hierarchical Information Extraction

I want to be able to analyze a single page or a sequence of pages from the concatenated dataframe.

Therefore, the code is designed to take in a defined page range.
As one cannot fully rely on the hierarchical level predicted by the layout segmentation model, I resort to taking the table of contents as a reference. The information from the TOC can be manually curated with limited effort. For this little showcase I manually added a column "TOC_hierarchy" encoding the hierarchy information from the TOC. I awarded a level of 20 for the highest hierarchy found and decreased the value for each level by 1. This is counter intuitive to the approach of the layout segmentation that identifies 0 as the highest hierarchy and might be reconsidered at a later stage.

Masterarbeit_HierarchieResorting > Hof > court_df_TOC.csv > data

```
,page,index,pred_text,class,H6_type,address,info,TOC_hierarchy
0,13,13_1,Hofstaat,H0,,,,20
1,13,13_2,"Seiner Oesterreichisch-kaiserlichen und königlichApostolischen Majestät.",H1,,,,19
2,13,13_3,Oberste Hofämter.,H1,,,,18
3,13,13_4,Erster Obersthofmeister.,H2,,,
```

I defined a function to be able to specify the page range. I enriched the information from the selected pages with regard to the type and hierarchical level of each element. I introduced columns with boolean values that discriminate between persons and organizational entities. Furthermore, within the organizational entities I made a discrimination between functions and institutions, with functions being the lowest hierarchical level (i.e. the last heading before one or more paragraphs) that normally assign specific roles to persons. Hierarchy was already added manually from the table of contents, but the TOC doesn't encompass all headings. A second indication of hierarchy are subsequent headings. If after one heading there immediately follows a second, the second is normally of a lower hierarchy. I tried to model this in an extra column "institutional_hierarchy", which starts with 1 for functions and increases by 1 for every immediately preceding heading.

In [1]:
```python
def build_section(start, stop):
    pages = list(range(start, stop + 1))
    #pages = [str(p) for p in pages]

    # import the dataframe that was enriched with hierarchy information from the TOC
    court_df = pd.read_csv('court_df_TOC.csv', index_col=0)

    # filter the dataframe for the pages that are within the specified range
    section_df = court_df[court_df["page"].isin(pages)].reset_index()

    # tries in boolean indexing to give structure to dataframe

    # all paragraphs or big paragraphs should be persons
    # I introduce a new column "is_person" that is true if the class is Paragraph
    # or BigParagraph
    section_df["is_person"] = np.where((section_df['class'] == "Paragraph") |
                                (section_df['class'] == "BigParagraph"), True, False)

    # all headings carry some information on organisational structure or hierarchy
    # I introduce a new column "is_orga" that is true if the class is H0, H1, H2, H3, H4, H5 
    section_df["is_orga"] = np.where((section_df['class'] == "H0") |
                                (section_df['class'] == "H1") |
                                (section_df['class'] == "H2") |
                                (section_df['class'] == "H3") |
                                (section_df['class'] == "H4") |
                                (section_df['class'] == "H5") |
```

```
                                (section_df['class'] == "H6"), True, False)

    # to determine whether a heading is a function or an institution, I need to look at the
    # sunsequent element. If the subsequent element is a paragraph, the heading is a function.
    section_df["is_function"] = section_df["class"]

    for i in range(len(section_df)):
        if section_df.loc[i, "is_function"] in type_heading and section_df.loc[
            i+1, "is_function"] in type_paragraph:
            section_df.loc[i, "is_function"] = True
        else:
            section_df.loc[i, "is_function"] = False

    # if a heading is a function, it is not an institution
    # therefore, I introduce a new column "is_institution" that is true if the heading is
    # an organisation but not a function
    section_df["is_institution"] = (section_df["is_function"] - section_df[
        "is_orga"]).astype("bool")

    # if two institutions are subsequent, the first is commonly of a higher hieracrchical leve
    # than the second

    # I introduce a new column "institutional_hierarchy" that carries the value of the
    # "is_institution" column of the subsequent element
    section_df["institutional_hierarchy"] = section_df[
        "is_institution"].tolist()[1:] + section_df["is_institution"].tolist()[:1]
    # I convert the column from bool to integer
    section_df["institutional_hierarchy"] = section_df[
        "institutional_hierarchy"].astype("int")
    # The code can handle up to three subsequent hierarchical entities. If there are more,
    # the value is set to 3. This simplification should be sufficient for a first analysis.
    section_df["institutional_hierarchy"] = (section_df[
        "is_institution"] + section_df["institutional_hierarchy"] + 1) * section_df[
            "is_institution"]
    # functions are theorganisational entity of the lowest hierarchical level and are set to 1
    section_df.loc[section_df["is_function"] == True, "institutional_hierarchy"] = 1

    return section_df
```

In a next step, which is coded as a separate function, I need to combine the hierarchical levels I have transcribed from the table of contents with those that were calculated by the rule of preceding headings. I unify them into one column but make use of the big value difference to discriminate between the two. I then ensure that institutional elements between those that have an TOC-assigned value are adapted accordingly.
In preparation of the visualization, I create a second dataframe "superiors_df" that juxtaposes each element with its hierarchically immediately preceding element.

[4]:
```
def redo_hierarchies(df):
    section_df = df

    # If the element has a hierarchy level assigned from the TOC, this level is copied to the
    # "institutional_hierarchy" column
    section_df.loc[section_df.TOC_hierarchy.isin(range(21)),
                   "institutional_hierarchy"] = section_df.loc[
```

```
                    section_df.TOC_hierarchy.isin(range(21)), "TOC_hierarchy"]

# I only work on the hierarchy column and therefore convert it to a list
re_hierarchy = section_df["institutional_hierarchy"].values.tolist()

# I determine the validated values in the hierarchy column. I define a lower cut-off at 9.
validated_vals = [x for x in set(re_hierarchy) if x>=9]
# I determine the indexes of the validated values in the list
validated_indexes = [i for i,x in enumerate(re_hierarchy) if x in validated_vals]
# The indexes of the validated values define range boundaries between which the hierarchy
# values need to be adjusted
ranges = [0] + validated_indexes + [len(re_hierarchy)]
# I determine the highest hierarchy value of the non-validated values
non_val_vals = [num for num in re_hierarchy if num not in validated_vals]
max_val = max(non_val_vals)

# for each range between two validated values, I adjust the non-validated values
# values in ranges need always be lower than any validated value
diff = 0
for i in range(len(validated_indexes)+1):
    for u in range(ranges[i], ranges[i+1]):
        # for the first value in the range, I set increment as the difference between
        # itself and the highest non-validated value
        # such,  that each of the values in the range decrease by one starting from the le
        # the first value in the range
        if re_hierarchy[u] in validated_vals:
            diff = re_hierarchy[u] - max_val
        else:
            re_hierarchy[u] += diff - 1

# I convert the list back to a pandas series and assign it to the dataframe to replace the
# old hierarchy values
section_df["institutional_hierarchy"] = re_hierarchy

levels = sorted(set(re_hierarchy), reverse=True)

# I want to create a new dataframe that only contains the name of the element and the
# name of the superior element in the hierarchy
superiors_df = section_df[["index", "pred_text", "institutional_hierarchy"]]
# I need to chain index value to "pred_text" field, to disambiguate functions with the sam
# name but from different institutions eg. "Director"
superiors_df['pred_text'] = superiors_df[['index','pred_text']].apply(
    lambda x : '{} {}'.format(x[0],x[1]), axis=1)

superiors_df = superiors_df.drop("index", axis=1)

# In preparation of the loop, I create a new column "superior" that is empty and save it
# to a list. I do the same with the hierarchy values.
superiors_df["superior"] = ""
superiors_df["superior"] = superiors_df["superior"].astype(str)
superiors = superiors_df["superior"].to_list()
hierarchy = superiors_df["institutional_hierarchy"].to_list()

# I loop through the hierarchy values from the highest to the lowest level. For each level
# I determine the elements that have this level assigned and then assign the name of the
# superior element to all elements that have a lower level and are subsequent to the eleme
for i in levels:
```

```
        sub_df = superiors_df.loc[superiors_df.institutional_hierarchy == i]
        sub_idx = superiors_df.loc[superiors_df.institutional_hierarchy == i].index.values.to
        # Starting from the highest level, I locate the name of the current superior element
        for u in sub_idx:
            superior = superiors_df.loc[u, "pred_text"]

            # I loop through the elements that have a lower level and are subsequent to the
            # current element. I assign the name of the superior element to the "superiors"
            # list for each of these elements at the respective index.
            for s in range(len(superiors_df)):
                if s > u:
                    if hierarchy[s] > i:
                        break
                    elif hierarchy[s] < i:
                        superiors[s] = superior


    # I assign the list of superiors to the dataframe
    superiors_df['institutional_hierarchy'] = hierarchy
    superiors_df['superior'] = superiors

    # The function returns two dataframes. The first contains the original dataframe with the
    # adjusted hierarchy values. The second contains the names of the elements and their
    # respective superiors in the hierarchy.
    return section_df, superiors_df
```

I defined another function that just executes the two preceding functions sequentially for the specified page range and then returns the two dataframes.

[5]:
```
def extract_hierarchy(start, stop):
        section_df = build_section(start, stop)
        section_df, superiors_df = redo_hierarchies(section_df)

        return section_df, superiors_df
```

Another function takes the superiors_df and creates a simple graph-like visualization.

[6]:
```
def draw_graph(superiors_df, start, stop):

    # The networkx library is used to create a graph from the superiors-dataframe
    # The library needs a list of tuples that contain the name of the element and the name of
    # the superior element. These tuples are called edges.
    edges = superiors_df[["pred_text", "superior"]]

    # The length of the names is limited to 40 characters to ensure that the graph is readable
    encode_list1 = edges["pred_text"].values.tolist()
    encode_list1 = [f'"{w[:40]}"' for w in encode_list1]
    edges["pred_text"] = encode_list1

    encode_list2 = edges["superior"].values.tolist()
    encode_list2 = [f'"{w[:40]}"' for w in encode_list2]
    edges["superior"] = encode_list2
```

```
    # The graph is created and saved as a png file
    orgchart=nx.from_pandas_edgelist(edges,
    source='superior', target='pred_text')

    p=nx.drawing.nx_pydot.to_pydot(orgchart)
    p.write_png(f'results/orgchart_{start}-{stop}.png')
```

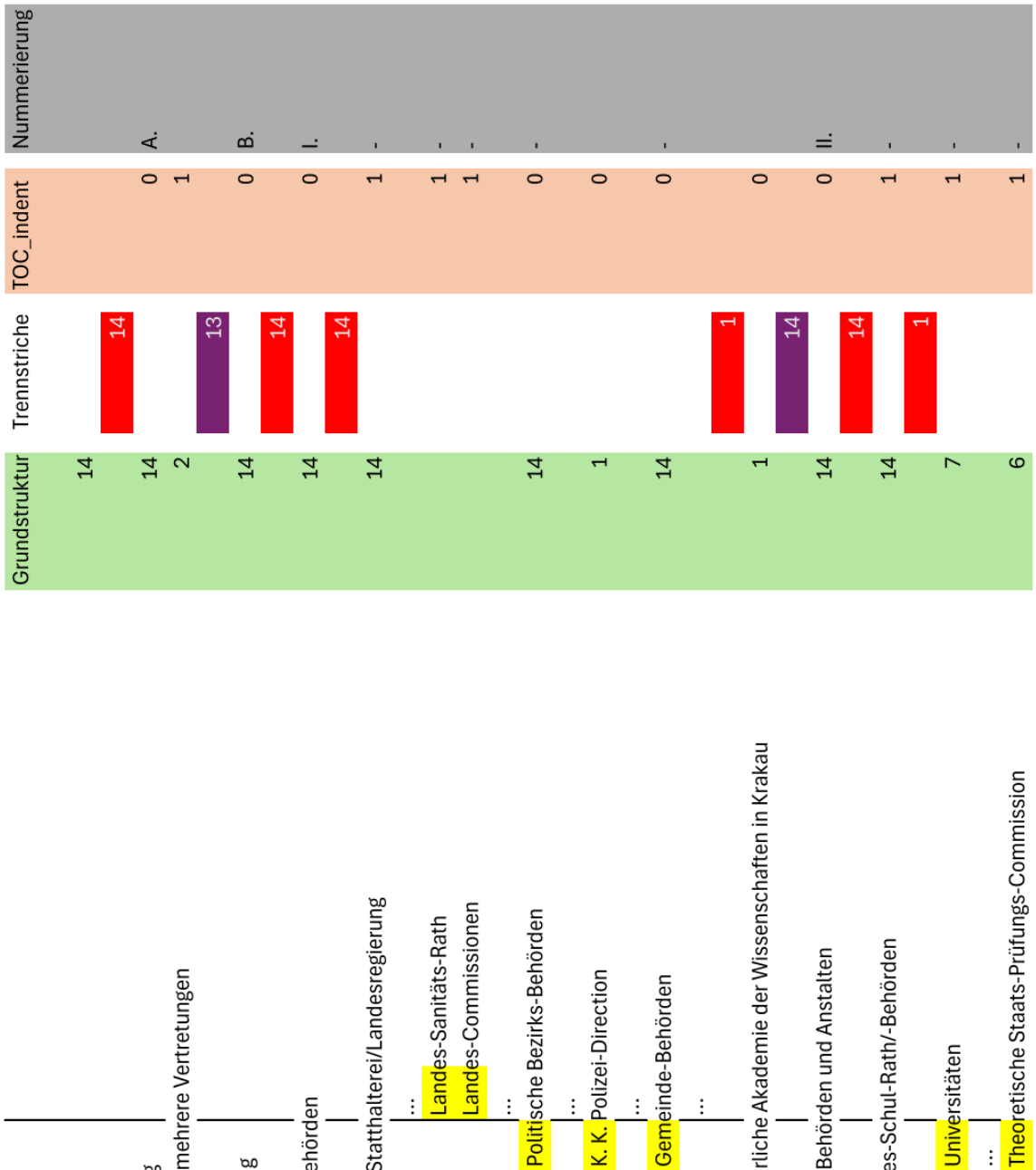
Now we can call the functions and export both the visualization as PNG and the detailed dataframe as CSV.

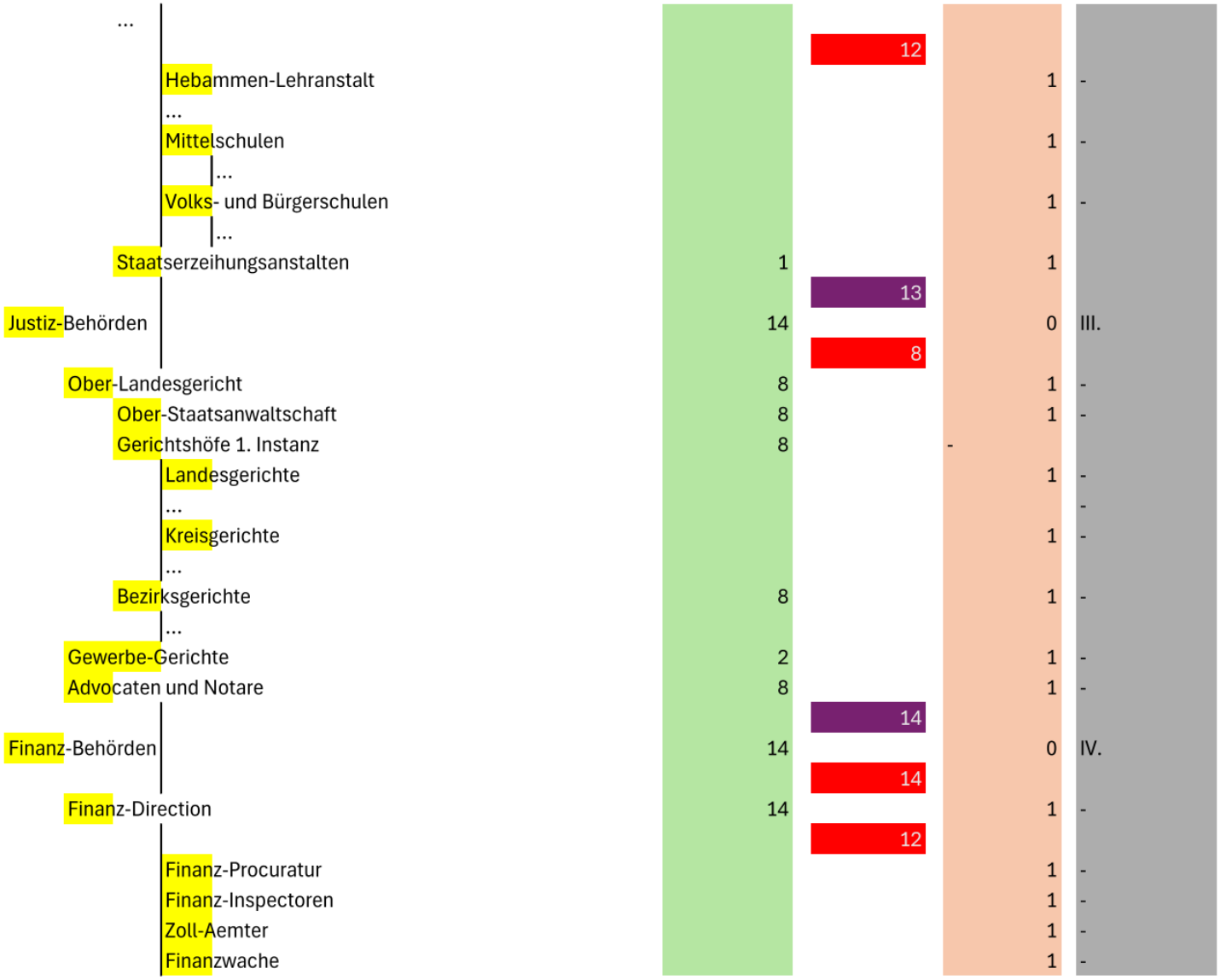
In [ ]:
```
# import pages 13-21 separately
start = 13
stop = 15


section_df, superiors_df = extract_hierarchy(start, stop)
draw_graph(superiors_df, start, stop)
section_df.to_csv(f'results/section_{start}-{stop}_analysis.csv')
```

References

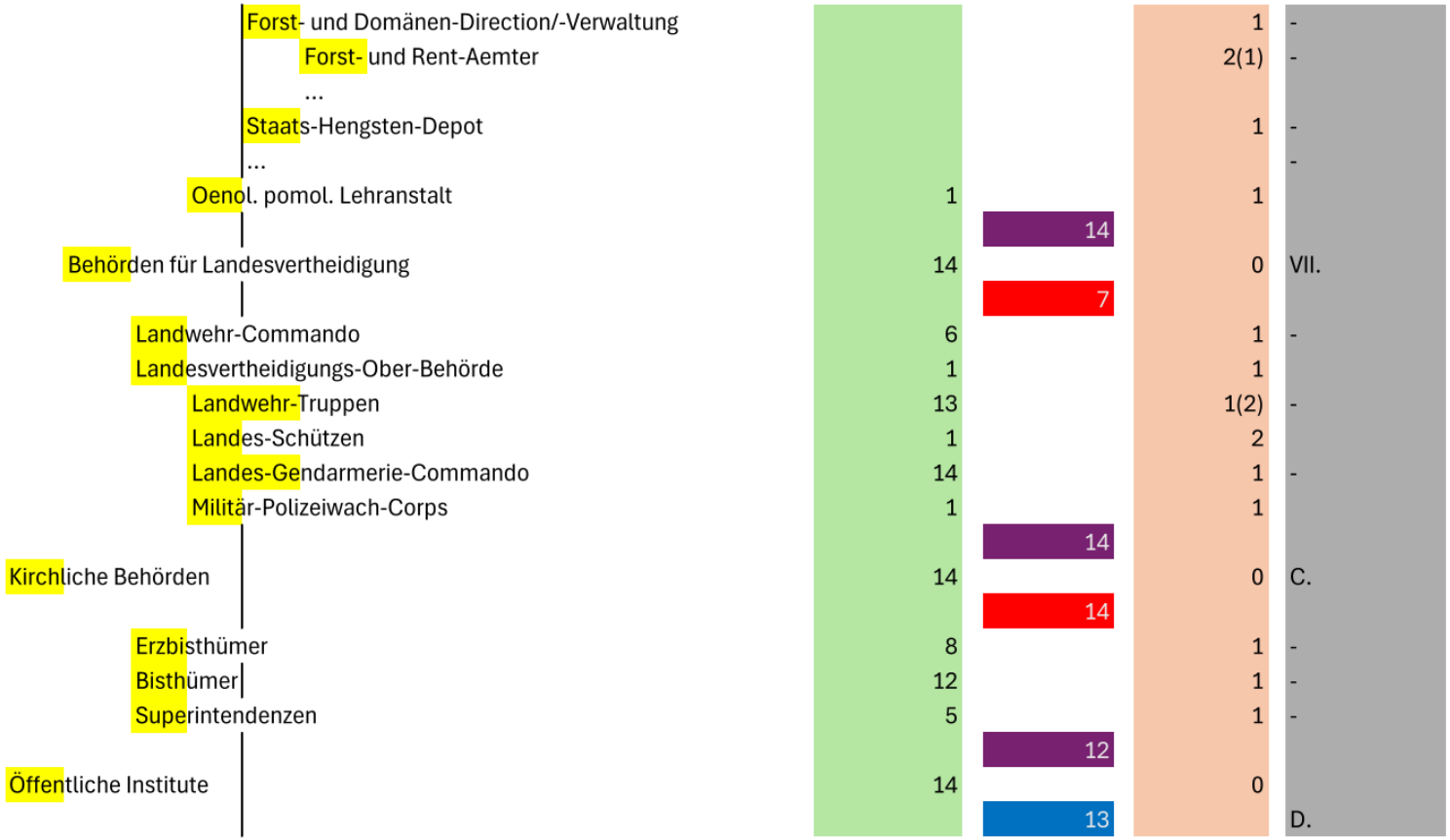
[1]: Fleischhacker, David, Wolfgang Goederle, and Roman Kern. "Improving OCR Quality in 19th Century Historical Documents Using a Combined Machine Learning Based Approach." January 15, 2024. http://arxiv.org/pdf/2401.07787.

10Annex 2: Assessment Table




...
12
Hebammen-Lehranstalt
1 -
...
Mittelschulen
1 -
...
Volks- und Bürgerschulen
1 -
...
Staatserzeihungsanstalten
1
1
13
Justiz-Behörden
14
0 III.
8
Ober-Landesgericht
8
1 -
Ober-Staatsanwaltschaft
8
1 -
Gerichtshöfe 1. Instanz
8
-
Landesgerichte
1 -
...
-
Kreisgerichte
1 -
...
Bezirksgerichte
8
1 -
...
Gewerbe-Gerichte
2
1 -
Advocaten und Notare
8
1 -
14
Finanz-Behörden
14
0 IV.
14
Finanz-Direction
14
1 -
12
Finanz-Procuratur
1 -
Finanz-Inspectoren
1 -
Zoll-Aemter
1 -
Finanzwache
1 -

Finanz-Bezirks-Direction			1	-
Landes-Haupt-Cassa			1	-
Landes-Filial-Cassa			1	-
Tax-Amts-Cassa			1	-
Steuer-Administration/Local-Commission			1	-
Verwaltungs-Organe für die directe Besteuerung			1	-
Haupt-Steuer-Aemter			1	-
Steuer-Aemter			1	-
Salinen-Ämter			1	-
Punzierungsstätten			1	-
Tabak-Fabriken			1	-
Lotto-Amt			1	-
...				
		1		
Wiener Börse	1		1	-
...				
		14		V.
Behörden für Handel und Verkehr	14		0	
Post-Direction			1	-
Post-Aemter			2(1)	-
Telegraphen-Direction			1	-
Telegraphen-Aemter			2(1)	-
Aich-Inspectorat			1	-
Aichämter			2(1)	-
Handels- und Gewerbekammer			1	-
...				-
		14		
Behörden für Landescultur und Bergwesen	14		0	VI.
Berg-Hauptmannschaft/-Behörden			1	-
...				


Forst- und Domänen-Direction/-Verwaltung
Forst- und Rent-Aemter
...
Staats-Hengsten-Depot
...
Oenol. pomol. Lehranstalt
Behörden für Landesvertheidigung
Landwehr-Commando
Landesvertheidigungs-Ober-Behörde
Landwehr-Truppen
Landes-Schützen
Landes-Gendarmerie-Commando
Militär-Polizeiwach-Corps
Kirchliche Behörden
Erzbisthümer
Bisthümer
Superintendenzen
Öffentliche Institute
1
14
14
7
6
1
13
1
14
1
14
14
8
12
5
12
14
13
1
2(1)
1
-
1
14
0
VII.
7
1
-
1
1(2)
-
2
1
-
1
14
0
C.
14
1
-
1
-
1
-
12
0
13
D.