Journal of Cultural Analytics

Cha, Javier. 2026. Fine-Tuning the Historian's Macroscope: Data Reuse and Medieval Korean Biographical Records in Neo4j. *Journal of Cultural Analytics*, 11(1): pp. 1–32. DOI: https://doi.org/10.22148/jca.1027

Fine-Tuning the Historian's Macroscope: Data Reuse and Medieval Korean Biographical Records in Neo4j

Javier Cha, History, University of Hong Kong, javiercha@hku.hk

This article explores the development and application of a historian's macroscope, a computational framework that enables multiscalar exploration of medieval Korean biographical records using Neo4j. While digitization has greatly expanded access to historical sources, existing methodologies struggle to integrate large, heterogeneous datasets while maintaining interpretive rigor and contextual specificity. This research demonstrates how a graph database model enhances historical inquiry by enabling the dynamic traversing of patronage and kinship networks. Unlike conventional network visualization tools, which often impose a rigid divide between individual detail and macrostructural patterns, Neo4j facilitates seamless transitions across different levels of analysis. This discussion illustrates how the macroscope helps historians uncover obscured relationships, refine analytical focus, and generate new research questions. More broadly, it highlights the need for methodologies in digital history that move beyond large-scale visualization to critically engage with the complexities of historical sources.

Introduction

Historians of premodern Korea enjoy unparalleled access to nearly every primary source from before 1900, owing to the South Korean government's sustained investment in cultural heritage digitization over the past twenty-seven years (Kim; Cha, "Digital/Humanities"; Cha, "Digital Korean Studies"). The Korean Historical Databases portal alone hosts more than one hundred collections,[1] and the raw data files underpinning these platforms are freely downloadable via the Open Data Portal.[2] Compared to their counterparts of Japan, China, and elsewhere, Korea specialists encounter few paywalls, institutional barriers, or copyright restrictions.

For those of us interested in the machine-assisted examination of historical dynamics, the unprecedented availability of digitized sources offers new opportunities but also presents its own set of challenges. My inquiry into patronage and marriage networks in medieval Korea, for example, cannot be pursued through relational databases (RDB), historical geographic information systems (HGIS), network visualization tools (e.g., Gephi or Cytoscape), or other standard digital historical methods. One obstacle lies in the sparse nature of the surviving sources, preserved only as fragmented clusters that resemble snowball samples. Expanding such research therefore requires close scrutiny of source reliability, careful contextualization, and attention to how one bundle of materials relates to another. These difficulties are compounded by the institutional design of digital repositories, which prioritize broad public access at the expense of the more specialized needs of professional historical inquiry.

In search of an adaptable framework, I turned to the notion of a macroscope—a set of computational methods or research environments that promises to reveal patterns and relationships in an extensive archival collection. The idea of the macroscope as a means of exploring large corpora predates the digital turn. In 1958, Philip Bagby envisions it as "an instrument which would ensure that the historian would see only the larger aspects of history and blind him to the individual details" (128). The aim, he argued, was to maintain "this higher level of abstraction [to] decipher the principal patterns of historical change" (Bagby 128). Information scientist Katy Börner defines it as a collection of algorithms, tools, plug-ins, and web services designed to "provide a 'vision of the whole,' helping us synthesize the related elements and detect patterns, trends, and outliers while granting access to myriad details" (60).

---

[1] National Institute of Korean History, *Han'guksa teit'ŏbeisŭ* [Korean Historical Databases], accessed 15 February 2025, https://db.history.go.kr/.

[2] Ministry of the Interior and Safety, *Konggong teit'ŏ p'otŏl* [Open Data Portal], accessed 15 February 2025, http://data.go.kr/.

Subsequently, the macroscope entered digital humanities discourse for its promise to enable researchers to explore corpora at both large and small scales. Instead of serving as "the opposite of the scientists' microscope" (Crymble 18), macroscopes function as a system of "different tools for exploring different scales" (Graham, Milligan, and Weingart xvi), allowing researchers to "observe what is at once too great, slow, or complex for the human eye and mind to notice and comprehend" (Börner 60). Building on this premise, Shawn Graham, Ian Milligan, and Scott Weingart introduce "the historian's macroscope" as a method for identifying patterns and constructing narratives from big data through "new vantage points and tools" (xvi). In a parallel development, Timothy Tangherlini proposes a macroscope for computational folkloristics, one that facilitates the exploration of textual corpora and multimodal content at varying levels of granularity while situating stories within their social and intertextual contexts (10, 18).

Despite these aspirations, macroscopes in digital humanities have yet to live up to their promise. Instead of allowing a researcher to glide across levels of analysis, they tend to reinforce the divide between isolated data points and structural patterns, inadvertently echoing Bagby's vision in which particulars fade in favor of laws and regularities. Graham, Milligan, and Weingart acknowledge this limitation, noting that macroscopes, by design, generate "text abstractions or data visualizations in lieu of direct images" (1). Yet most current implementations offer only modest ways of adjusting scales dynamically. A historian's macroscope worth the name ought to permit fluid movement across layers of analysis, especially the middle ground where local contexts intersect with broader structures, as Tangherlini shows in computational folkloristics. Such flexibility would make it possible to study populations, social groups, and regional variations while still setting them against wider temporal and geographic horizons. This was, after all, the distinctive strength of the Annales school, with its advocacy for shifting the historian's temporal vantage between *longue durée*, *conjoncture*, and *événements*, and the geographic scope between the expanse of the Mediterranean world and the close scrutiny of a single French village. Macroscopes, if properly designed, should offer a comparable mobility: digital readings and rereadings of machine-readable sources that make it possible to bridge macro-level structures and microhistorical case studies, uncovering relationships otherwise obscured by scale, complexity, fragmentation, or simple loss.

If macroscopes promise new vantage points, they also reveal how difficult it remains to align large-scale digital collections with the needs of historical research. Digital historians of East Asia and elsewhere have encountered the limits of current

methodologies, particularly in how institutional repositories are structured, queried, and interpreted, and in how common software tools are designed for broad use rather than project-specific inquiries. As an alternative, I propose that historians create personal digital libraries and bespoke research environments tailored to their projects rather than rely on macro-level "distant readings" of centralized repositories. Medieval Korea provides a particularly revealing case because its surviving sources are fragmentary and diverse, making it an ideal context for testing how dispersed information might be organized and interlinked. To that end, I show how building a project-specific environment with Neo4j, a graph database management system, makes it possible to bring such fragments into dialogue with one another. This experiment emphasizes contextualization and authentication, improving how heterogeneous records can be linked, organized, and queried. After working with general-purpose macroscopes such as Gephi and Cytoscape, and recognizing their limitations, I outline a workflow that uses Python to draw selected subsets of data from larger archives and structure them in Neo4j. The result is not only a technical solution but also a research strategy—one that enables historians to refine their analytical scope and open up new avenues of inquiry.

Lost in Data: The Unfulfilled Potential of Macroscopes

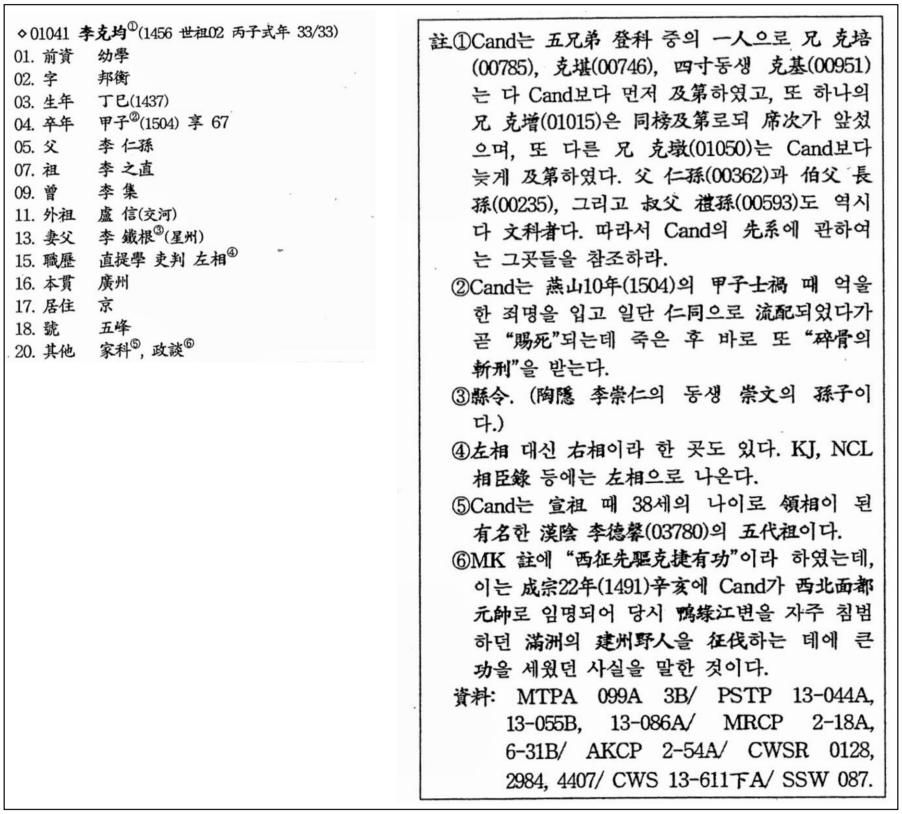
In 1967, Edward Wagner and Song June-ho secured a Ford Foundation grant to launch the Munkwa Project, an ambitious effort to construct a database of approximately 14,600 Korean men who passed the *munkwa* civil service examinations between 1393 and 1894.[3] Working at breakneck speed on punch card technology despite the absence of standardized encoding for Chinese and Korean characters, they presented their preliminary findings at the 1970 annual meeting of the Association for Asian Studies (Wagner, "Computer Study") and followed up with a series of influential publications. Their work immediately reshaped the field of Korean history, bringing to light for the first time the Pareto distribution of degree-producing clans, the dominance of the capital region, and the eighteenth-century rise of northerners (Wagner, "Northern Provinces"). Yet Wagner's vision extended far beyond generating aggregate values. He sought to use the exam roster as a proxy for understanding Korea's "unusual degree of continuity in its elite structure from the traditional into the modern period" (Wagner, "Project Description" 6), a theme that preoccupied him for the rest of his career. To pursue this objective, Wagner and Song attempted to link examination rosters with the

[3] The *munkwa* was the highest-level state examination in Chosŏn (1392–1910), serving as the primary pathway to civil officialdom in the central government. Candidates underwent rigorous testing on their knowledge of the Confucian classics, history, literature, and statecraft.

vast collection of early modern family genealogies, but the challenges of reconciling disparate sources proved insurmountable. The project remains unfinished long after both scholars passed away.

The unfulfilled promise of the Munkwa Project offers valuable lessons for digital historians. One of the persistent challenges of digital historical research is the tendency to privilege large-scale analysis at the expense of other essential aspects of a historian's workflow. When digital historians construct macroscopes, the elevated view relegates critical evaluation to the margins—or buries it in an overlooked "notes" column in a data table. This approach makes it difficult to corroborate findings across many kinds of evidence and constrains the historian's ability to shift perspectives for contextualization. Wagner and Song relied heavily on the notes column when attempting to link the civil service examination roster with extended family information from genealogies (**Figure 1**). While this method allowed them to append records, it also obscured the process of authentication and validation, complicating efforts to trace original sources, correct discrepancies, or reconcile conflicting details. This issue is not unique to Korean historical scholarship. Ian Milligan has shown how the digital edition of the *Toronto Star* omitted an entire microfilm reel containing entries on Franklin Roosevelt's visit to Toronto, which was an oversight that went unnoticed until his discovery (*Transformation* 24). Digital editions inherit the limitations of their source materials, and without rigorous scrutiny, such gaps and distortions can affect the quality of research and shape scholarly conclusions in unintended ways. Treating source criticism as secondary compromises the reliability of historical interpretation.

Another enduring challenge for digital historians is the integration of diverse forms of evidence. Many tools, methods, and research environments are designed to provide a bird's-eye view of a single database, usually through statistical modeling, spatial visualization, or network analysis. Yet, as Adam Crymble reminds us, history is fundamentally "an argument-driven, evidence-based answer to a question about the past" (18). Historical inquiry depends on synthesizing information from multiple sources, not on privileging a single collection, event, or type of material. For instance, historians studying state and society relations in Korea draw not only on official court accounts but also on local gazetteers, litigation records, and diaries. The pursuit of *histoire problème* may extend further still, incorporating non-documentary and intangible evidence such as ceramics, metallurgy, printing technologies, or climate data. Because most macroscopes are optimized for an individual collection, however, data-driven historical research tends to operate within these constraints.



◇01041 李克均①(1456 世祖02 丙子式年 33/33)
01. 前資 幼學
02. 字 邦衡
03. 生年 丁巳(1437)
04. 卒年 甲子⑫(1504) 享 67
05. 父 李 仁孫
07. 祖 李 之直
09. 曾 李 集
11. 外祖 盧 信(交河)
13. 妻父 李 鐵根③(星州)
15. 職歷 直提學 吏判 左相④
16. 本貫 廣州
17. 居住 京
18. 號 五峰
20. 其他 家科⑤, 政談⑥

註①Cand는 五兄弟 登科 중의 一人으로 兄 克培(00785), 克堪(00746), 四寸동생 克基(00951)는 다 Cand보다 먼저 及第하였고, 또 하나의 兄 克增(01015)은 同榜及第로되 席次가 앞섰으며, 또 다른 兄 克墩(01050)는 Cand보다 늦게 及第하였다. 父 仁孫(00362)과 伯父 長孫(00235), 그리고 叔父 禮孫(00593)도 역시 다 文科者다. 따라서 Cand의 先系에 관하여는 그곳들을 참조하라.
②Cand는 燕山10年(1504)의 甲子士禍 때 억울한 죄명을 입고 일단 仁同으로 流配되었다가 곧 "賜死"되는데 죽은 후 바로 또 "碎骨의 斬刑"을 받는다.
③縣令. (陶隱 李崇仁의 동생 崇文의 孫子이다.)
④左相 대신 右相이라 한 곳도 있다. KJ, NCL 相臣錄 등에는 左相으로 나온다.
⑤Cand는 宣祖 때 38세의 나이로 領相이 된 有名한 漢陰 李德馨(03780)의 五代祖이다.
⑥MK 註에 "西征先驅克捷有功"이라 하였는데, 이는 成宗22年(1491)辛亥에 Cand가 西北面都元帥로 임명되어 당시 鴨綠江변을 자주 침범하던 滿洲의 建州野人을 征伐하는 데에 큰 功을 세웠던 사실을 말한 것이다.
資料: MTPA 099A 3B/ PSTP 13-044A, 13-055B, 13-086A/ MRCP 2-18A, 6-31B/ AKCP 2-54A/ CWSR 0128, 2984, 4407/ CWS 13-611下A/ SSW 087.

Figure 1: The entry for Yi Kŭkkyun (1437–1504) in the Wagner–Song Munkwa database shows a detailed notes section.[4]

[4] The following is an English translation of the Wagner–Song Munkwa entry number 01041 for Yi Kŭkkyun (1437–1504). The entry makes extensive use of "Others" and "Notes" fields, supplemented by a heavily acronym-laden list of sources that form its evidentiary basis. While the kinship details are useful for individual searches, the database schema is not designed for multi-level or longitudinal analysis.
Left:
01041 Yi Kŭkkyun (1456, second reign year of King Sejo, pyŏngja year, ranked 33rd out of 33 successful candidates)
01. Previous position: student
02. Courtesy name: Panghyŏng
03. Year of birth: chŏngsa (1437)
04. Year of death: kapcha (2) (1504), aged 67
05. Father: Yi Inson
07. Paternal grandfather: Yi Chijik
09. Paternal great-grandfather: Yi Chip
11. Maternal grandfather: No Sin (Kyoha)
13. Father-in-law: Yi Ch'ŏlgŭn (3) (Sŏngju)
15. Career records: Second Deputy Director, Minister of Personnel, Second State Councilor
16. Ancestral seat: Kwangju
17. Residence: capital
18. Sobriquet: Obong
20. Others: other family members with examination degrees(5), political intrigues and stories

This pattern is evident in recent studies: Lee Sangkuk and Lee Wonjae's work on medieval marriage practices relies entirely on the 1476 edition of the Andong Kwŏn genealogy, while Lee Sangkuk and Park Jong Hee's analysis of aristocratic survival strategies draws on the 1565 edition of the Munhwa Yu genealogy. Hŏ Su's conceptual history of the early twentieth century runs topic models on the intellectual magazine *Opening of the World* (*Kyebyŏk*). Beyond Korea, similar cases include Ian Miller's study of rebellion in early modern China based on the Qing court annals as well as Robert Nelson's digital reading of the American Civil War using the Richmond newspaper *Daily Dispatch*.[5] Yet the discipline of history has long pieced together evidence from multiple sources, weighed contradictions, and situated details in both narrower and broader contexts.[6] Historians are trained to move fluidly across sources and scales, but our digital environments encourage a panoramic view fixed on a single corpus. Unless this gap is closed, digital history risks producing data exercises stripped of the synthesis and source criticism that give the discipline its interpretive force.

Right:

Notes

1. The candidate was one of five brothers who passed the civil service examination. His older brothers Kŭkpae (00785), Kŭkkam (00746), and his cousin Kŭkki (00951) all passed the exam before him. Another older brother, Kŭkchûng (01015), took the examination in the same year but ranked higher, while yet another older brother, Kŭkton (01050), passed a later examination. His father, Inson (00362), along with elder paternal uncle Changson (00235) and younger paternal uncle Yeson (00593), also held examination degrees. To trace the candidate's lineage further, consult their respective entries.
2. In the tenth reign year of Yŏnsan'gun (1504), the candidate was falsely accused during the Kapcha Purge and exiled to Indong, where he was forced to commit suicide by poison sent from the king. After his death, his body was ordered to be dismembered.
3. A local magistrate and grandson of Toŭn Yi Sungin's younger brother, Sungmun.
4. Some records list him as Third State Councillor rather than Second. In KJ, NCL, the *Roster of State Councilors*, etc., he appears as Second State Councillor.
5. The candidate was the fifth-generation ancestor of the famous Hanŭm Yi Tŏkhyŏng (03780), who became Chief State Councillor at the age of 38 during King Sŏnjo's reign.
6. A note in MK states that he "led the vanguard of the western campaign, achieving triumph and distinction," referring to an event in 1491 when the candidate was appointed Supreme Commander of the Northwestern Frontier. His leadership resulted in a major military victory, including the conquest of the Jianzhou Jurchens from Manchuria, who frequently crossed the Amnok River to raid the region.

Sources: MTPA 099A 3B/ PSTP 13-044A, 13-055B, 13-086A/ MRCP 2-18A, 6-31B/ AKCP 2-54A/ CWSR 0128, 2984, 4407/ CWS 13-611 lower A volume/ SSW 087.

[5] Robert K. Nelson, "Mining the Dispatch," accessed 15 February 2025, https://dsl.richmond.edu/dispatch/about.

[6] Ian Miller's study of rebellion, banditry, and crime in Qing China is a model of careful method. He openly acknowledges the limits of using latent Dirichlet allocation topic modeling on the *Veritable Records* (*Qing shilu*), the official court annals of the Qing dynasty 629–635). Instead of diving straight into statistical analysis, Miller begins with a thoughtful reflection on the role of "secret memorials"—confidential reports that alerted the emperor and his closest advisers to violent incidents across the empire, among other urgent concerns. By foregrounding this context, Miller shows an acute awareness that while the *Veritable Records* largely document major offenses, they also provide one of the most comprehensive windows onto patterns of violence in Qing society and the conditions under which unrest could flare into upheaval and rebellion.

A third challenge concerns the changing nature of primary sources, in both digital and non-digital forms, and their entanglement with institutional and individual research agendas. Again, the Munkwa Project serves as a vivid example. The examination data for the period before 1600 relies on seventeenth-century reconstructions produced in the aftermath of a devastating war with Japan and subsequent Manchu incursions, which destroyed or dispersed many original records (Park 178–9). Further complications arise from the coexistence of official and commercial editions, which diverge in significant ways, particularly in biographical and kinship details (Park 193–201). In addition, modern digital editions of these sources bear complex histories, shaped as much by institutional priorities as by personality clashes. In the 1980s, tensions between Edward Wagner at Harvard University and Yi Sŏngmu at the Academy of Korean Studies (AKS) led to the creation of independent AKS exam databases, which today are more widely used than the Wagner–Song edition. Since then, individual scholars have reorganized, corrected, and adapted these rosters, producing personalized datasets, including those by Lee Jaeok, Park Hyun Soon, and myself. Likewise, Korean cultural heritage databases reflect institutional agendas, competition over funding, human error, and the underpaid labor of data entry personnel—all of which shape the digital resources available to historians.

Finally, digital historians face the daunting challenge of managing vast historical records. When Wagner and Song decided to supplement the examination rosters with genealogical data (Wagner, "Korean Chokpo"), they did not fully grasp the magnitude of the task. Early modern Korean genealogies are exponentially larger than examination rosters. *The Grand Genealogy of Myriad Surnames* (*Mansŏng taedongbo*), for example, documents 136,000 individuals across 361 choronyms, while *Augmented Genealogy of the Royal Family* (*Sŏnwŏn sokpo*) records 560,000 members of the dynastic clan's extended lineage. And these are just two among hundreds of genealogies available to specialists of medieval and early modern Korea, each containing overlapping information as well as discrepancies in biographical details and kinship connections (Park 193–208). Because genealogies document far more individuals than the 14,600 civil examination records in their original database, the scale of their undertaking surpassed their initial expectations. At this level of complexity, manual verification becomes practically impossible.

In response to the abovementioned challenges, I developed the Medieval Yangban Project, a Neo4j-based macroscope designed to connect and reuse heterogeneous datasets in ways that preserve the methodological rigor of traditional historical research while also opening new possibilities for reading primary sources in a robust digital environment. Instead of defaulting to "distant readings" of institutional archival collections, this project integrates biographical records, genealogies, and official

documents to investigate specific questions about the formation of medieval Korea's *yangban* aristocracy. The Medieval Yangban Project offers an environment that aligns with core historical principles—corroboration, contextualization, and the careful calibration of macro- and micro-historical perspectives—and also enables the historian to piece together disparate evidence, trace relationships, and revise interpretations as new material comes to light. The following section outlines the data sources incorporated into the Medieval Yangban Project and explains why managing and reusing research data through Neo4j offers distinct advantages over existing digital methods.

The Medieval Yangban Project: Data Sources

In an ongoing study, I examine the formation of medieval Korea's *yangban* aristocracy through the lens of patronage and marriage networks (Cha, "Medieval Patrimonialism"). Monarchs, regents, and others in the upper echelons recruited and elevated loyal followers. Provincial newcomers who secured political sponsorship, in turn, sought to transform fleeting opportunity into lasting privilege for their descendants. A handful managed it. Most did not. And only a very few were able to maintain their footing in the capital for more than three generations. Over time, the select few among the latter consolidated into hereditary *yangban* descent groups that continued to shape governance into the early modern era. In this social figuration, one of the most important formal institutions was the state examination system, which, on paper, offered talented individuals a meritocratic pathway into central officialdom. In practice, however, success depended as much on political sponsorship as on ability.

To make sense of these evolving social figurations in medieval Korea, the Medieval Yangban Project turns to Neo4j to link and reuse a wide range of biographical data assembled from fragmented and scattered historical records. This machine-assisted approach builds on the promises and limitations of Wagner and Song's Munkwa Project, which pioneered the digitization of the civil service examination rosters and genealogical materials but struggled to bring disparate sources into a coherent analytical framework. Expanding on that effort, the Medieval Yangban Project brings together biographical, kinship, and official records alongside other documentary sources, enabling a multiscalar view of Korea's premodern elite. The Medieval Yangban Project also recognizes that not all sources are created equal. How these materials were compiled, digitized, and released matters, and these historical processes need to be taken seriously and examined critically. This work goes beyond identifying the sources themselves in order to include reconciling inconsistencies across datasets, linking identifiers used by different data providers, and grappling with the uneven survival of historical records.

The core biographical data for the Medieval Yangban Project is drawn from *The Comprehensive Information System for Korean Historical Figures* (*Han'guk yŏktae inmul chonghap sisŭt'em)*, an online platform maintained by AKS.[7] Developed between 2005 and 2011 through multiple rounds of institutional funding from the National Information Society Agency (Academy of Korean Studies, "Han'guk yŏktae inmul"), this system consolidates three major repositories: (1) records of 108,392 degree holders of the civil, military, or technical state examinations, each tagged with a unique identifier and an exam type code indicating the qualification attained (**Figures 2** and **3**); (2) biographical data of 27,020 notable personalities; and (3) official registers of those who held government posts (**Figure 2**). These records, available as XML files, can be accessed through the South Korean government's Open Data Portal.

Filename	Contents
20141023_역대인물_과거파일_EXM.zip	exam degree holders in 106,132 .xml files
20141022_역대인물_인물사전_PPL.zip	notable individuals in 27,043 .xml files
20141022_역대인물_관인정보_GOV.zip	officials in 48,561 .xml files

Figure 2: XML files for *The Comprehensive Information System for Korean Historical Figures* downloaded via the Open Data Portal contain the raw data and metadata of exam degree holders, officials, and other notable individuals.

AKS Exam Type Code	Meaning
EXM_KM	Koryŏ *munkwa* (higher civil)
EXM_KS	Koryŏ *saengwŏn* (lesser civil)
EXM_MN	Chosŏn *munkwa* (higher civil service)
EXM_S1	Chosŏn *saengwŏn* (understanding of classics)
EXM_S2	Chosŏn *chinsa* (literary composition)
EXM_MU	Chosŏn *mukwa* (military ability)
EXM_Z1	language translation
EXM_Z2	medical knowledge and skills
EXM_Z3	astrological knowledge
EXM_Z4	legal expertise
EXM_Z5	miscellaneous talent recruitment

Figure 3: This table lists the exam type codes used by the Academy of Korean Studies to identify Koryŏ and Chosŏn civil, military, and technical examinations.

[7] Academy of Korean Studies, *Han'guk yŏktae inmul chonghap sisŭt'em* [The Comprehensive Information System for Korean Historical Figures], accessed 15 February 2025, http://people.aks.ac.kr/.

During preprocessing, I constructed a master table cataloging each historical figure's multiple identifiers, as assigned by the different project teams responsible for digitization (**Figure 4**). One identifier, the encykorea key, links an individual to their entry in *The Grand Encyclopedia of Korean Culture* (*Han'guk minjok munhwa taepaekkwa sajŏn*), first published in 1991 as a twenty-seven-volume print edition and later expanded into a CD-ROM version in 2001. Another system, the Universal Content Identifier (UCI), was introduced by AKS to standardize references across datasets and assist with disambiguation (Academy of Korean Studies, "Yŏktae inmul UCI"). Its uneven application, however, means that the Medieval Yangban Project treats UCI as one of several entity markers instead of a definitive linking mechanism. Reconciling and merging records from these different sources requires a systematic approach—a task well suited to Neo4j.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	name	birthyear	deathyear	ancestralse	choronym	choronym2	office_type1	office_type2	office_type3	father	gfather	biog_id	UCI	exam_id	encykorea	period
2	1	眞空	855	937				종교	불교인		金確宗		PPL_5COa	G002+AKS-KHF_12C9(		E0054582	고려 전기
3	2	李惠言	858	938	碧珍	碧珍 李	碧珍 李	기타	지방세력가				PPL_5COa	G002+AKS-KHF_13C77		E0046250	고려 전기
4	3	雁嚴	862	930	慶州	慶州 金	慶州 金	종교	불교인				PPL_5COa	G002+AKS-KHF_12C5E		E0036419	고려 전기
5	4	允多	864	945				종교	불교인				PPL_5COa	G002+AKS-KHF_12C72		E0042262	고려 전기
6	5	崔彦撝	868	944	慶州	慶州 崔	慶州 崔	문신	문신				PPL_5COa	G002+AKS-KHF_13CD!		E0057458	고려 전기
7	6	慶甫	869	948				종교	불교인		金益良		PPL_5COa	G002+AKS-KHF_12ACE		E0002430	고려 전기
8	7	璟幽	869	958				종교	불교인		金容		PPL_5COa	G002+AKS-KHF_12CC2		E0055234	고려 전기
9	8	忠湛	869	940				종교	불교인				PPL_5COa	G002+AKS-KHF_12CD/		E0058114	고려 전기
10	9	靈照	870	947				종교	불교인				PPL_5COa	G002+AKS-KHF_12C6(		E0037668	고려 전기
11	10	慶猷	871	921				종교	불교인				PPL_5COa	G002+AKS-KHF_12ACE		E0002730	고려 전기
9398	10900	張紅株						충효열	기타				PPL_6JOb	G002+AKS-KHF_23C7A	5D64DB3C		조선 중기
9399	10901	鄭文榮妻						문학	시·시조인				PPL_6JOb	G002+AKS-KHF_24C815	BB38C601		조선 중기
9400	10902	林碧堂 金氏			義城	義城 林	義城 林	문학	시·시조인				PPL_6JOb	G002+AKS-KHF_25C784	BCBDB2F(		조선 중기
9401	10903	李泰陽						유생					PPL_6JOb	G002+AKS-KHF_13C774	D0DCC59		조선 중기
9402	10904	曺典言						여관(女官)					PPL_6JOb	G002+AKS-KHF_23C870	C804C5B8		조선 중기
9403	10905	文益周	1535	1605	南平	南平 文	南平 文	문신	문신				PPL_6JOc	G002+AKS-KHF_13BB3	E0019631		조선 중기

Figure 4: The Medieval Yangban main table shows multiple identifiers, including UCI and encykorea.

Not all records from the aforementioned datasets serve the aims of the Medieval Yangban Project. Only materials pertaining to *yangban* lineages, state examination candidates, and officeholders are included. The project examines the period from 900 to 1600, spanning the Koryŏ dynasty (918–1392) and the first two centuries of Chosŏn (1392–1910). The survival of sources from this era was significantly shaped by two major events. The first was the dynastic change from Koryŏ to Chosŏn in 1392, which brought sweeping reorganization and reinterpretation of official records. The second was the Imjin War (1592–1598), when many pre-1600 texts were lost or destroyed, leaving a thin documentary base for the study of both Koryŏ and early Chosŏn. The scale of this loss can be seen in *A Comprehensive Compilation of Korean Collected Works* (*Han'guk munjip ch'onggan*), which compiles the writings of 1,259 premodern authors. The five centuries corresponding to Koryŏ (c. 901–1400) yield only 21,193 Sinitic characters, while for Chosŏn scholars the figure jumps nearly

twenty-five-fold to 540,639 for the period 1401 to 1900. The imprint of the Imjin War is also evident: before 1600, only 90,030 characters survive, or about 17 percent of the Chosŏn total, compared to 450,609 written after 1601 (**Figure 5**).[8]

Index Year Range	Total Characters (*cha* 字)
901 to 1400	21,193
1401 to 1600	90,030
1601 to 1900	450,609
1401 to 1900 (approximately Chosŏn)	540,639
901 to 1900 (approximately Koryŏ and Chosŏn combined)	561,832

Figure 5: Total character count in *A Comprehensive Compilation of Korean Collected Works* shows the variation of data within indexed year ranges.

A similar pattern emerges in government records. The 50 million Sinitic characters that comprise *The Annals of the Chosŏn Dynasty* (*Chosŏn wangjo sillok*), covering 1392 to 1910, remains fully intact and is available as an annotated set of XML documents. Compiled after a king's death, these records were based on *The Diary of the Royal Secretariat* (*Sŭngjŏngwŏn ilgi*), a more detailed daily chronicle maintained by royal scribes. However, the pre-1623 volumes of *The Diary of the Royal Secretariat* were lost during the Imjin War and a subsequent rebellion. As a result, historians studying the latter half of Chosŏn have access to over 242 million characters in its digital edition, while those examining earlier periods are confined to the more condensed and selective accounts preserved in *The Annals of the Chosŏn Dynasty*.

From the institutional biographical databases, the Medieval Yangban Project narrows its focus to 9,402 individuals whose index years fall between 900 and 1600. The index year system follows conventions developed by the China Biographical Database Project (CBDB), which assigns a reference year based on known birth or death dates

[8] These figures require further refinement, as collections from earlier centuries sometimes include later writings added by descendants or admirers as a gesture of respect. For instance, the surviving edition of *The Collected Works of Sŏng Hyŏn* (1439–1504) is based on an 1842 reprint of an eighteenth-century manuscript copy because the original sixteenth-century publication was lost. This edition contains several prefaces written by his descendants in the nineteenth century. The current character count system treats these nineteenth-century additions as part of Sŏng Hyŏn's fifteenth-century writings, given their negligible proportion relative to the main text. Moving forward, I plan to improve this system to produce a more precise character count by century.

(China Biographical Database).[9] If a birth year is available, the fifty-ninth birthday is used; if not, or if the recorded year of death falls before that, then the year of death is assigned instead. The dataset itself is organized with Python's ElementTree parser and the Ablebits plugin for Excel, and it includes fields such as name, birth and death years, place of origin, and cross-referenced identifiers. Out of caution, the dataset of government officials has not been incorporated at this stage, given concerns about data reliability and the heavy skew toward post-1600 figures.

The dataset on examination degree holders provides additional insight into intergenerational mobility and kinship networks. Among the 108,392 total degrees conferred, only 19,557 fall within the period between 958 and 1608, corresponding to Koryŏ's first state examination and the death of King Sŏnjo (r. 1567–1608) of Chosŏn. Of these, 15,743 were civil degrees, 3,131 were military, and 683 were in technical fields (**Figure 6**). The examination records also include genealogical data, listing an individual's father, grandfather, great-grandfather, father-in-law, and maternal grandfather—according to Confucian patrilineal principles. This allows for the reconstruction of kinship ties using an edge list of agnatic and affinal relationships. The Medieval Yangban Project repurposes 28,087 records of ego-to-son and ego-to-son-in-law relationships, matched to examination degree holders through various identifiers. A Python web scraper is used to map these examination records to the UCI system (**Figure 7**).

AKS Exam Code	Total Entries
KM	1,169
KS	394
MN	4,853
MU	3,131
S1	4,824
S2	4,503
Z1	220
Z2	175
Z3	71
Z4	60
Z5	157

Figure 6: This table lists the distribution of examination degree holders from the Academy of Korean Studies database included in the Medieval Yangban Project.

[9] Paul Vierthaler has kindly brought to my attention that CBDB's index year rule was recently updated. In a future release of the Medieval Yangban Project dataset, I plan to incorporate this revision, with both the old and new rules applied.

In [ ]:
```
from urllib.request import urlopen
import urllib
from bs4 import BeautifulSoup
import pandas as pd
```

In [ ]:
```
# 고려문과
html = urlopen('http://people.aks.ac.kr/front/dirSer/exm/exmKingExmList.aks?classCode=KM&className=%EA%B3%A0%EB%A0%A4%EB%AC
soup = BeautifulSoup(html, 'html.parser')
king_href_list = []

kings = soup.find_all('td', { 'headers': 'king_name' })
for king in kings:
    king_href = king.find('a', href=True)
    king_href = king_href['href']
    king_href = urllib.parse.quote_plus(king_href, safe='/?=&')
    king_href = 'http://people.aks.ac.kr' + king_href
    print(king_href)
    king_href_list.append(king_href)
```

In [ ]:
```
exam_href_list = []

for king_href in king_href_list:
    html = urlopen(king_href)
    soup = BeautifulSoup(html, 'html.parser')
    exams = soup.find_all('td', { 'headers': 'king_name' })
    for exam in exams:
        exam_href = exam.find('a', href=True)
        exam_href = exam_href['href']
        exam_href = urllib.parse.quote_plus(exam_href, safe='/?=&')
        exam_href = 'http://people.aks.ac.kr' + exam_href
        print(exam_href)
        exam_href_list.append(exam_href)
```

In [ ]:
```
cand_href_list = []

for exam_href in exam_href_list:
    html = urlopen(exam_href)
    soup = BeautifulSoup(html, 'html.parser')
    candidates = soup.find_all('td', { 'headers': 'fullname' })
    for candidate in candidates:
        cand_href = candidate.find('a', href=True)
        cand_href = cand_href['href']
        cand_href = urllib.parse.quote_plus(cand_href, safe='/?=&')
        cand_href = 'http://people.aks.ac.kr' + cand_href
        print(cand_href)
        cand_href_list.append(cand_href)
```

In [ ]:
```
name_UCI = []

for cand_href in cand_href_list:
    html = urlopen(cand_href)
    soup = BeautifulSoup(html, 'html.parser')
    name = soup.find('div', {'id': 'contentBody_title' })
    uci = soup.find('a', {'id': 'uci'})
    item = [cand_href, name.text.strip(), uci.text]
    print(item)
    name_UCI.append(item)
```

In [ ]:
```
df =pd.DataFrame(name_UCI)
df.to_csv('km_uci.csv', encoding='utf-8', sep='\t')
```

Figure 7: Python code retrieves EXM and universal content identifiers from *The Comprehensive Information System for Korean Historical Figures*.

To complicate the picture, the Medieval Yangban Project augments its core biographical data with genealogies, which are especially valuable for tracking kinship networks across multiple generations. They also serve as a check on the exam rosters, at times confirming the details they contain, at other times identifying inconsistencies. One key source is *A Genealogy of Myriad Clans* (*Man'gabo*), a large-scale compilation

produced in the late nineteenth century by the Haenam Yun, one of early modern Korea's wealthiest *yangban* families. Digitized by AKS, this work contains records of 116,451 individuals and 270 descent groups. A complementary source, the aforementioned *Grand Genealogy of Myriad Surnames*, first published in 1933 and digitized in 2014 by Ha Yŏnghwi and Paek Kwangyŏl as part of the Lineage Network Information System, extends this coverage.[10] Together, these genealogical sources make it possible to trace kinship ties among examination candidates, officeholders, and other notable historical figures.

The Medieval Yangban Project also incorporates Lee Jaeok's critical scholarly edition of the state examination database. Developed as part of his PhD dissertation, Lee's customized and manually verified dataset enhances the official AKS database and resolves many of its inconsistencies.[11] His edition assigns unique identifiers to individuals and uses suffixes to indicate kinship ties, such as father, grandfather, son, and brother. The dataset comprises both an earlier version and a later revision, and the Medieval Yangban Project references both. Legacy identifiers for degree holders begin with prefixes like "PMn," "PSa," and "PMu," while kinship links are marked with suffixes such as "A1," "A2," "B1," and "C1" (**Figure 8**). In his later revisions, Lee also aligned these records with genealogical data found in *A Genealogy of Myriad Clans*. In the next section, I show how I build on his meticulous work by linking individuals across multiple sources, resolving discrepancies, and reconstructing historical networks with the highest possible level of accuracy and reliability.

Data Reuse and Master Data Management in Neo4j

The Medieval Yangban Project is the result of seventeen years of experimental research on medieval Korean history using computational methods. Initially, I sought to merge Wagner and Song's original objectives with Börner's "vision of the whole" and Bagby's "higher level of abstraction" by applying network visualization with off-the-shelf software such as Gephi and Cytoscape. In 2008, however, the Korean Open Government License had yet to be enacted, and only Dongbang Media offered a commercial web edition of the Wagner–Song Munkwa Project, based on its original 2002 CD-ROM release. Due to pre-Unicode character encoding, Dongbang Media's digital preparation substituted image files for rare Sinitic characters. I spent several years manually resolving these encoding problems, correcting character errors and restoring missing entries one by one.

---

[10] Ha Yŏnghwi and Paek Kwangyŏl, "Chokpo kŏmsaekki," accessed 1 February 2025, http://lnis.kr/.

[11] Lee Jaeok, "Kwagŏ hapkyŏkcha chŏngbo tijit'ŏl ak'aibŭ," accessed 1 February 2025, https://dh.aks.ac.kr/~sonamu5/wiki/index.php.

	nid	full_name	name	chi_nan
1	PMu_1453_000001_A1	김덕로(金德老)	김덕로	金德老
2	PMu_1453_000002_A1	오영로(吳寧老)	오영로	吳寧老
3	PMu_1453_000007_A1	강속(姜涑)	강속	姜涑
4	PMu_1453_000008_A1	한양(韓讓)	한양	韓讓
5	PMu_1453_000009_A1	최지(崔池)	최지	崔池
6	PMu_1453_000010_A1	이언(李堰)	이언	李堰
7	PMu_1453_000011_A1	김흥인(金興仁)	김흥인	金興仁
8	PMu_1453_000012_A1	신유정(申惟精)	신유정	申惟精
9	PMu_1453_000013_A1	강흘(姜仡)	강흘	姜仡
10	PMu_1453_000014_A1	권시좌(權時佐)	권시좌	權時佐
11	PMu_1453_000015_A1	전유도(全由道)	전유도	全由道
12	PMu_1453_000019_A1	이작(李灼)	이작	李灼
13	PMu_1453_000020_A1	인인경(印仁敬)	인인경	印仁敬
14	PMu_1453_000021_A1	김중해(金仲海)	김중해	金仲海
15	PMu_1453_000022_A1	황균(黃均)	황균	黃均

Figure 8: Lee Jaeok's exam type and kinship identification system shows legacy identifiers and kinship links among other genealogical data.

In 2010 and 2011, my participation in an Institute for Advanced Topics in the Digital Humanities workshop on "Networks and Network Analysis for the Humanities" provided both the opportunity and motivation to preprocess a preliminary dataset of examination candidates who received degrees between 1393 and 1469, along with their kin, into Cytoscape for visualization.[12] In a dataset containing 3,793 edges and 4,057 nodes, Cytoscape's force-directed layout algorithm revealed a giant component with 2,040 edges and 1,936 nodes, alongside dozens of subnetworks (see **Figure 9**). Further visualization attempts in later years, using datasets covering different time periods and incorporating custom datasets created by other researchers, yielded similar results. **Figure 10** illustrates the Wagner–Song data for 1502–1609 in Gephi, while **Figure 12** presents a panorama of the "whole" network, a structure dubbed the "Eye of Munkwa," constructed using Lee Jaeok's Chosŏn civil examination data for 1393–1894, encompassing 47,293 nodes and 49,808 edges.

[12] See Networks and Network Analysis for the Humanities: An NEH Institute for Advanced Topics in Digital Humanities, Institute for Pure and Applied Mathematics, University of California–Los Angeles, 15–27 August 2010, https://www.ipam.ucla.edu/programs/summer-schools/networks-and-network-analysis-for-the-humanities-an-neh-institute-for-advanced-topics-in-digital-humanities/.

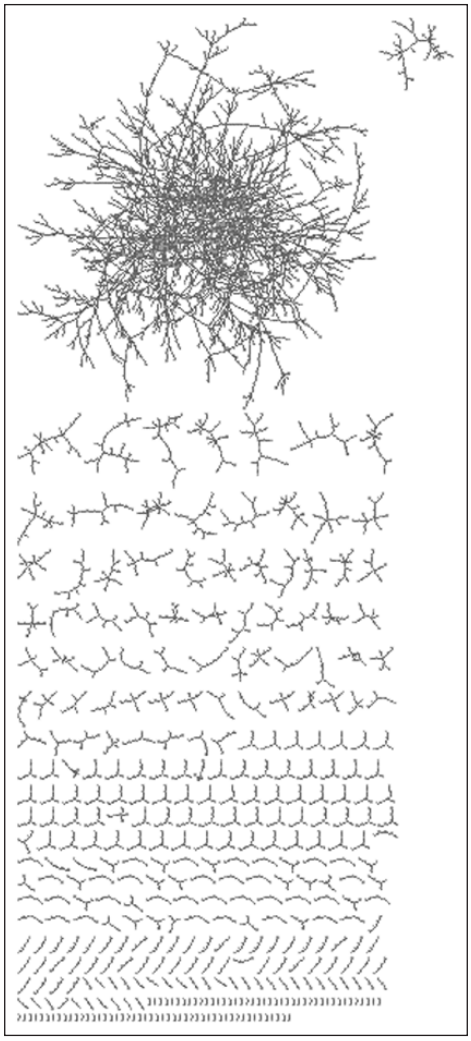


Figure 9: The kinship network of Chosŏn civil service examination degree holders (1393–1469) is visualized in Cytoscape and consists of 4,057 nodes and 3,793 edges.

Although Gephi and Cytoscape made it possible to "discover" the giant component among Chosŏn examination candidates linked by agnatic and affinal ties, these tools have proven unwieldy as effective historical macroscopes. Their limitations become apparent in their inability to implement Bagby's proposed shallow depth of field or Graham, Milligan, and Weingart's data abstractions. Gephi's capabilities are largely restricted to static network visualizations and statistical summaries of various centrality metrics, which provide little analytical value for genealogical datasets that do not conform to scale-free typologies, as genealogies are inherently tree structures. While Cytoscape allows for adjacent selection of specific nodes and edges and enables the extraction of subnetworks through filtering, its capacity to integrate

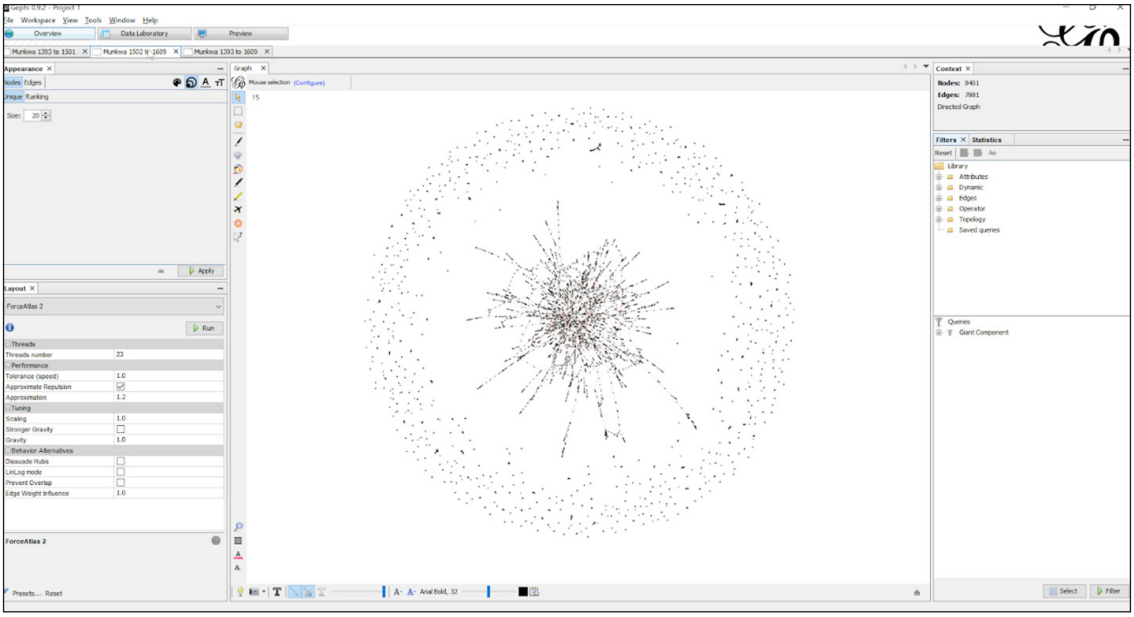


Figure 10: Wagner–Song Munkwa kinship data is visualized in Gephi and covers the years 1502–1609.

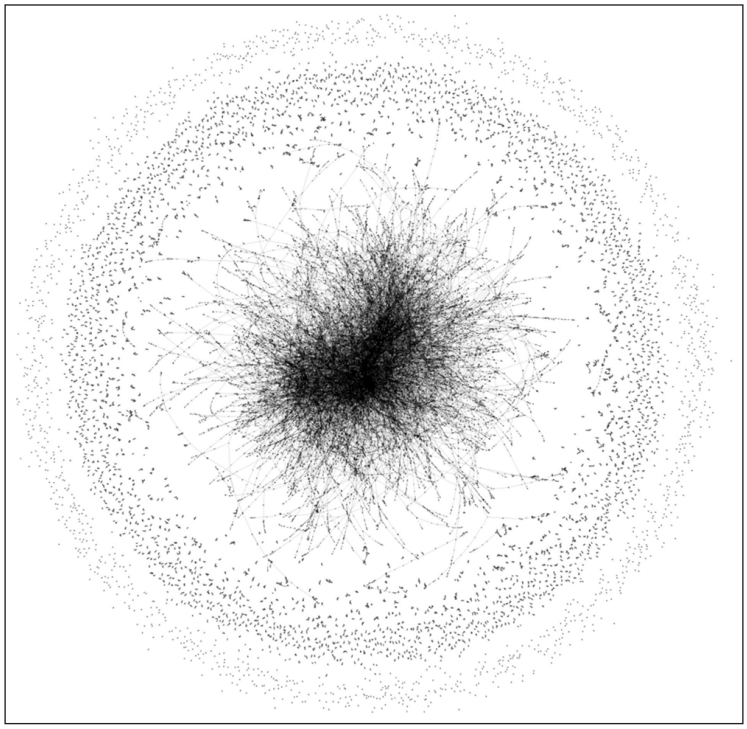


Figure 11: The kinship network of Chosŏn civil service examination degree holders, 1393–1894, is visualized in Gephi using Lee Jaeok's data (47,293 nodes and 49,808 edges), dubbed the "Eye of Munkwa."

and cross-reference multiple sources remains inadequate (**Figure 12**). The persistent challenges of volume and heterogeneity, which have affected digital Korean studies since the Munkwa Project, demand a more adaptable methodological approach.

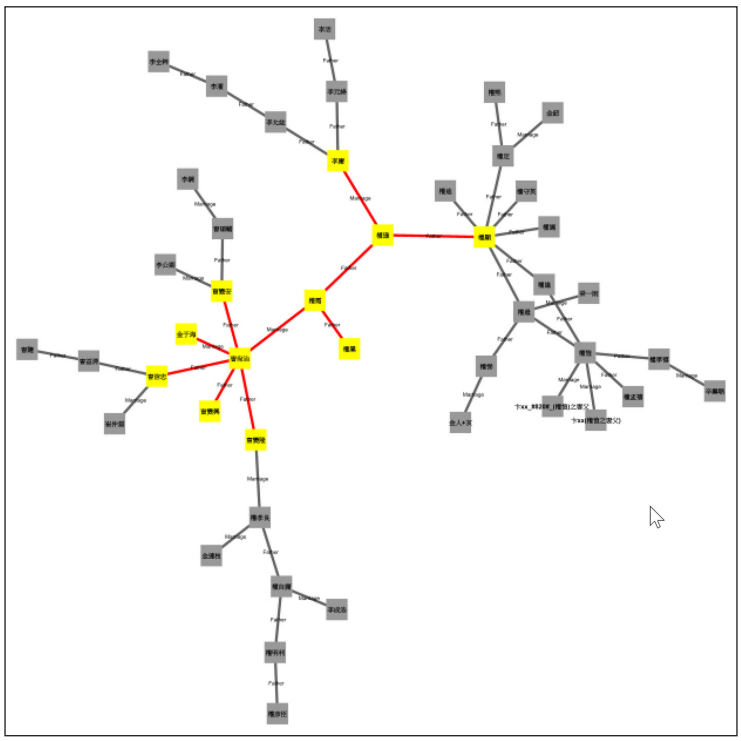


Figure 12: This figure demonstrates Cytoscape's adjacent node and edge selection feature.

The Medieval Yangban Project uses Neo4j as a central hub for integrating state examination rosters, biographical records, genealogies, and curated datasets contributed by digital historians of premodern Korea. In business applications, this approach is a well-established use case for graph database technology known as master data management.[13] My approach adapts this industry practice for historical research with a workflow that extracts, models, structures, and analyzes subsets of data specifically geared toward understanding processes of medieval aristocratic formation. Instead of interacting directly with large institutional repositories, my macroscope collects targeted datasets via Python scripts and imports them into Neo4j. In this environment, records undergo disambiguation, entity resolution, and graph-based querying to uncover patterns, outliers, and overlooked relationships.

The first step in this process is data extraction, which varies according to how the data sources are formatted. AKS provides biographical and examination data in XML, but the

[13] "Use Cases: Master Data Management," Neo4j, https://neo4j.com/use-cases/master-data-management/.

```
In [110]: for index, row in lnis_munhyong.iterrows():
              nodes = lnis_nodes[(lnis_nodes['name'] == row['name']) & (lnis_nodes['본관'] == row['본관'])]
              for index, row in nodes.iterrows():
                  lnis_id = int(row['lnis_id'])

                  # Degree 1: Ego appears in source
                  edges_son = lnis_edges_son[lnis_edges_son['source'] == lnis_id]
                  if not edges_son.empty:
                      for index, row in edges_son.iterrows():
                          print(row['source'], "\t", row['target'], "\t", row['relationship'])

                          # Degree 2: Ego's sources' sources
                          edges_son = lnis_edges_son[lnis_edges_son['target'] == row['source']]
                          for index, row in edges_son.iterrows():
                              print(row['source'], "\t", row['target'], "\t", row['relationship'])

                              # Degree 3: Ego's sources' sources' sources
                              edges_son = lnis_edges_son[lnis_edges_son['target'] == row['source']]
                              for index, row in edges_son.iterrows():
                                  print(row['source'], "\t", row['target'], "\t", row['relationship'])

                  # Degree 1: Ego appears in target
                  edges_son = lnis_edges_son[lnis_edges_son['target'] == lnis_id]
                  if not edges_son.empty:
                      for index, row in edges_son.iterrows():
                          print(row['source'], "\t", row['target'], "\t", row['relationship'])

                          # Degree 2: Ego's targets' targets
                          edges_son = lnis_edges_son[lnis_edges_son['source'] == row['target']]
                          for index, row in edges_son.iterrows():
                              print(row['source'], "\t", row['target'], "\t", row['relationship'])

                              # Degree 3: Ego's targets' targets' targets
                              edges_son = lnis_edges_son[lnis_edges_son['source'] == row['target']]
                              for index, row in edges_son.iterrows():
                                  print(row['source'], "\t", row['target'], "\t", row['relationship'])
```

Figure 13: This Python script extracts three generations of genealogical data above and below each ego.

UCI for each entry is available only on its online platform, which requires web scraping. Lee Jaeok's scholarly editions of examination and genealogical data are published in Excel, whereas the Lineage Network Information System dataset is available in JSON. Filtering tabular data by dynastic period, such as early Koryŏ, late Koryŏ, or mid-Chosŏn, or by index years, birth years, or death years facilitates the selection of relevant records. Using this method, I isolated 35 percent of biographical entries and 18 percent of conferred degrees that fall in the range of 900 to 1600. Genealogical records, however, pose greater challenges. Most entries contain only given names and choronyms, making it necessary to run a Python script with nested loops to identify kinship ties within three generations of each notable individual or examination candidate (**Figure 13**).

Once the data are extracted, the next step is to bring them into Neo4j in bulk, where I label nodes, establish relationships, merge duplicates, and resolve ambiguities. The Medieval Yangban Project uses Neo4j's labeling feature to represent people, titles, and sources, while the relationships record social, kinship, semantic, and bibliographic connections. Every individual is labeled as a "Person," but those who appear in the AKS exam database carry an additional tag that specifies their exam type. Lee Jaeok's dataset and genealogical sources are treated in a similar fashion so that their provenance is always clear. Inevitably, the import process produces duplicate records and overlapping relationships, which makes it crucial to have a method for unifying them. Yet this is more than a matter of routine de-duplication. It also demands thinking carefully about the very different ways relational and graph databases handle entities and linkage.

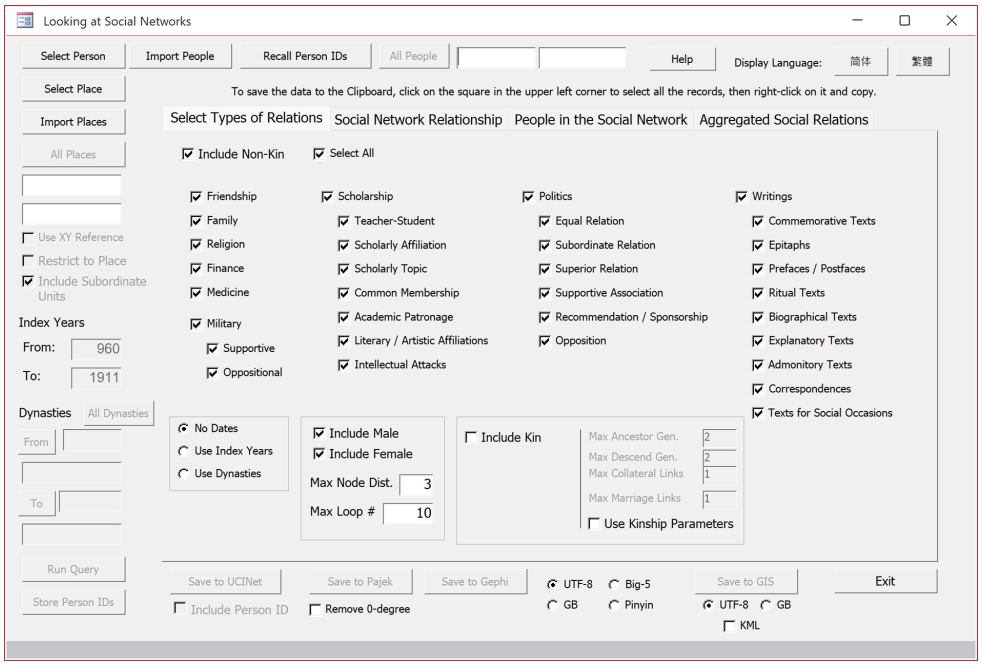


Figure 14: China Biographical Database's social network lookup applet iterates through the relational database and extracts the relevant social relations data as an edge list.

Neo4j-based management of heterogeneous records takes advantage of features not available in relational databases or XML, such as index-free adjacency, node labels, and longitudinal graph search. These capabilities allow historians to carry out machine-assisted readings and uncover relationships that might otherwise remain hidden. Neo4j's native ability to link and reuse curated datasets in a graph environment also creates space to ask new questions and refine ongoing ones using the Cypher query language. Most importantly, this graph-based approach is optimized for multiscalar exploration of both broad patterns and focused clusters of individuals, events, or concepts. Neo4j is not a panacea, of course, but when thoughtfully designed and implemented, a macroscope built on this system can lead to a more nuanced and comprehensive understanding of historical dynamics.

The Medieval Yangban Project stores all nodes and edges as graph data using Neo4j's native support for index-free adjacency. This makes it possible to begin with a telephoto view of an individual examination candidate or political faction and then zoom out to a wide-angle perspective of adjacent entities and regions without breaking continuity. The distinction becomes clear when compared to non-graph based solutions, such as the social network lookup applet for the China Biographical Database built in Visual Basic for Microsoft Access (**Figure 14**). That system requires the user to select an individual by unique identifier and trigger a series of SQL queries to retrieve relational connections through loops, with results exported to UCINet,

Pajek, or Gephi for visualization. By contrast, the Medieval Yangban Project performs such queries only once, during preprocessing and import, and embeds kinship and social ties directly into the database. As a result, tracing the extended network of, for example, the fifteenth-century scholar Kim Chongjik (1431–1492) does not require runtime SQL outer joins; rather, his kinship and political ties are already baked into the Neo4j database. Examples like this point to the larger promise of the project: to show how individual careers and family strategies, once pieced together, help illuminate the broader processes of aristocratic formation in medieval Korea (**Figure 15**).

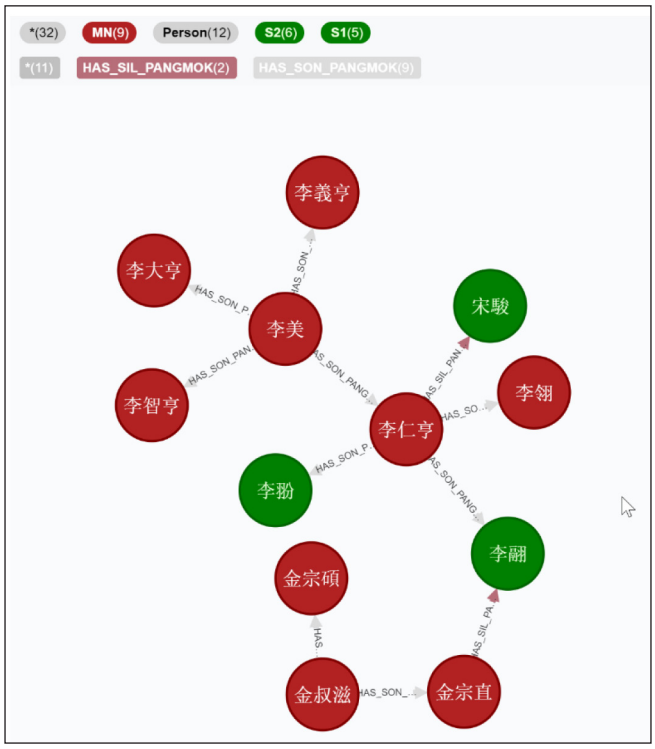


Figure 15: This example shows Kim Chongjik's (1431–1492) extended kinship network in the Neo4j browser.

That said, Neo4j is not a full substitute for relational databases. One of the main trade-offs of a graph database is the lack of third normal form and the absence of primary and foreign keys, both of which make it harder to safeguard data integrity.[14] Take duplicate

[14] Third normal form (3NF) is a database design principle that minimizes redundancy by ensuring non-key attributes depend only on the primary key. In relational databases, 3NF helps preserve data integrity by storing each piece of information in a single location. Rather than primary and foreign keys, Neo4j prioritizes rapid traversal across interconnected nodes, a procedure that offers greater flexibility but also risks compromising data integrity. For this reason, the Medieval Yangban Project requires a structured approach to detecting duplicates and merging them, a task that a relational database would ordinarily regulate.

nodes, for example. Cypher can flag potential duplicates with the WITH clause and collect() subquery, but it does not offer an easy way to merge them. For that, I turned to the Awesome Procedures on Cypher (APOC) extension, which adds a set of powerful refactoring tools, including apoc.refactor.mergeNodes(). This function lets the researcher decide whether to discard, overwrite, or combine properties into arrays when merging. I chose the combine mode, which preserves all property values—a decision that later proved especially useful for disambiguating records and catching anomalies (**Figure 16**).

Figure 16: The screenshots show the entries for Sŏ Kŏjŏng (1420–1488) before and after the deduplication process performed using apoc.refactor.mergeNodes().

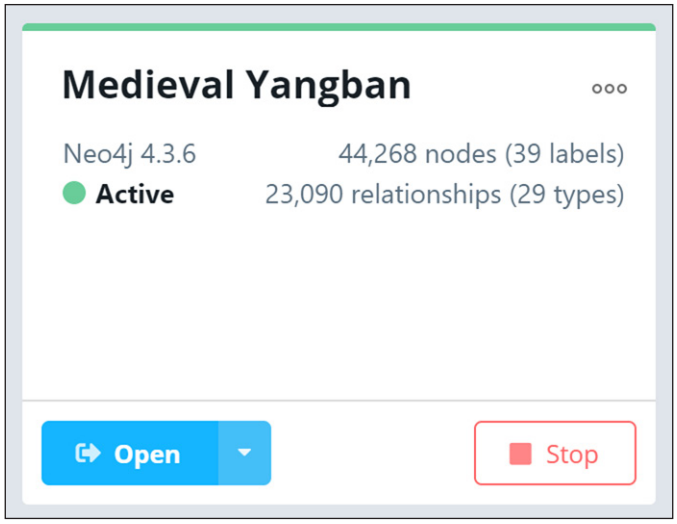


Figure 17: At the time of this article's publication, the Medieval Yangban Project macroscope in Neo4j comprises 44,268 nodes, 39 labels, and 23,090 relationships of 29 types.

To push Neo4j's capabilities further, I added new datasets tailored to additional research questions. These include 1,097 merit subjects who received investiture between 1327 and 1589, 370 collected works, and 4,487 individuals referenced in the *Annals of the Chosŏn Dynasty*. The merit subject dataset is especially useful for countering the common but misleading impression shaped by the historiography that this institution mattered only in the fifteenth century. The collected works dataset, by contrast, helps identify figures whose writings have survived. As present, the database contains 44,268 nodes with 39 labels—about twenty of which function as categories rather than entities—and 23,090 relationships of 29 types (**Figure 17**).

The use of multiple node labels and category nodes is particularly significant. Neo4j functions primarily as a graph data management system, and not a visualization or analysis tool, so labels provide a way to generalize connected components. This strategy aligns with recent innovations in computational humanities, particularly network-based models developed to make sense of large-scale historical and cultural datasets. In their work on Timothy Tangherlini's Danish folklore corpus, for example, his co-author James Abello and others propose modular decomposition to break down giant components similar to the ones shown in **Figures 9, 10**, and **11**, which are colloquially referred to as "hairballs" (Abello et al.). The Medieval Yangban Project adopts a similar strategy using Neo4j, incorporating not only patronage and kinship ties but also officeholding patterns, textual attributions, and merit subject investiture into a single framework, one that supports movement between highly abstracted overviews and more finely decomposed views of *yangban* networks.

With the Neo4j macroscope in place, I returned to Edward Wagner's classic research on the hereditary links of merit subjects across generations. In a 1974 study, Wagner challenged the common belief that the sixteenth-century reformer Cho Kwangjo (1482–1519) was a political outsider. He showed instead that Cho belonged firmly to the capital-based elite, tracing his lineage back to Cho On (1347–1417), who had received merit subject investiture for supporting the regime change of 1392 (Wagner, *Literati Purges* 80). Cho's case is illuminating, but one example alone can only take us so far in understanding the broader sociopolitical structures of late medieval Korea. To test Neo4j's potential, I turned to a question that historiography has largely passed over. Using a Cypher query, I identified all instances where the merit subjects of 1469 and 1589 were connected within two node distances (**Figure 18**). The reasoning here is simple: in Korean historiography, the merit subjects of 1469 have received sustained attention, while those of 1589 remain obscure. The clarity of the output owed much to the use of node labels and category nodes for merit subjects and collected works, which made it possible to generate an abstracted view of their interconnections.

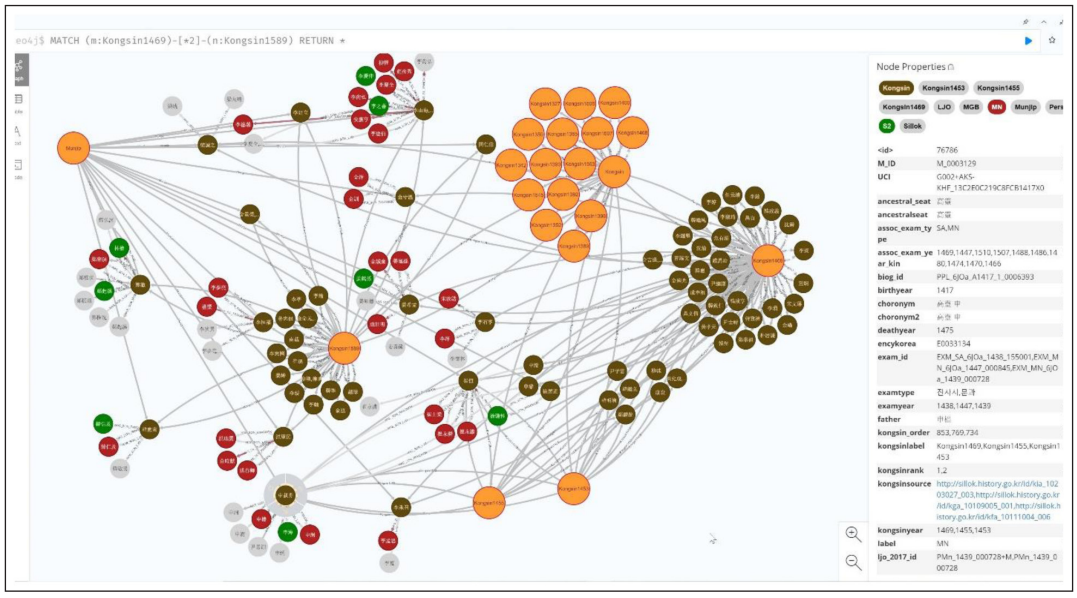


Figure 18: The screenshot shows all the links between merit subjects of 1469 and 1589 spanning two node distances, decomposed using category nodes to produce a simplified view.

Among the various experimental queries conducted on the Medieval Yangban Project, some of the most significant findings emerged through an APOC procedure called apoc.path.subgraphAll(). This procedure traverses a network from a specified set

of starting nodes and expands all subgraphs within a defined range. **Figure 19** shows Neo4j returning all merit subjects with a military background, up to three degrees of separation. The results suggest that most either had no known descendants or were purged shortly after their investiture. By contrast, the equivalent query for civil officials (**Figure 20**) reveals a dense network of intergenerational and affinal connections, implying a stable and steady coalescence of some of them into a self-perpetuating hereditary clusters.

The apoc.path.subgraphAll() procedure makes it possible to move from the abstracted view in **Figure 18** to a decomposed view of the merit subjects of 1469 and 1589, revealing details that would otherwise remain hidden. In **Figure 20**, for example, multiple chains of kinship ties stretch across numerous nodes. One connection in particular stood out: the long chain linking Sin Sukchu (1417–1475) and Hong Sŏngmin (1536–1594). To look more closely, I ran a targeted Cypher query focused on this relationship (**Figure 21**). What emerges is that density here does not necessarily signal centrality, but it does reliability. The father–son tie between Sin Chang (1382–1433) and Sin Sukchu, for instance, is corroborated by multiple sources, including the AKS examination database, Lee Jaeok's research edition, and *A Genealogy of Myriad Clans*. By contrast, the affinal link between Sin Songju (1429–1503) and Hong Yundŏk (dates unknown) appears only in the AKS examination data—a reminder that even seemingly straightforward connections may need careful verification.

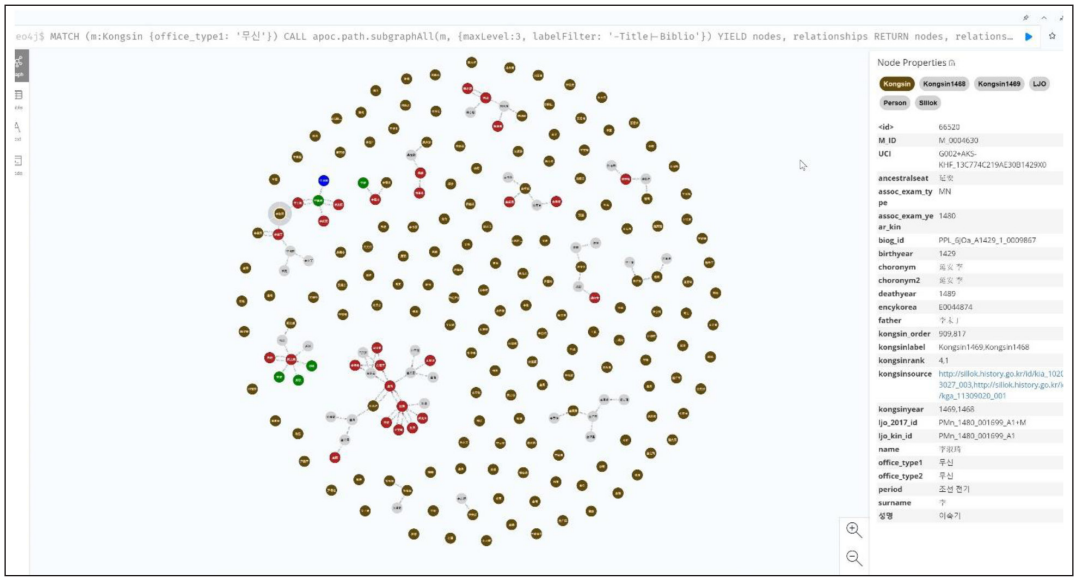


Figure 19: This figure shows merit subjects with a military background up to three degrees of separation. The results suggest that most either left no known descendants or were purged shortly after their investiture.

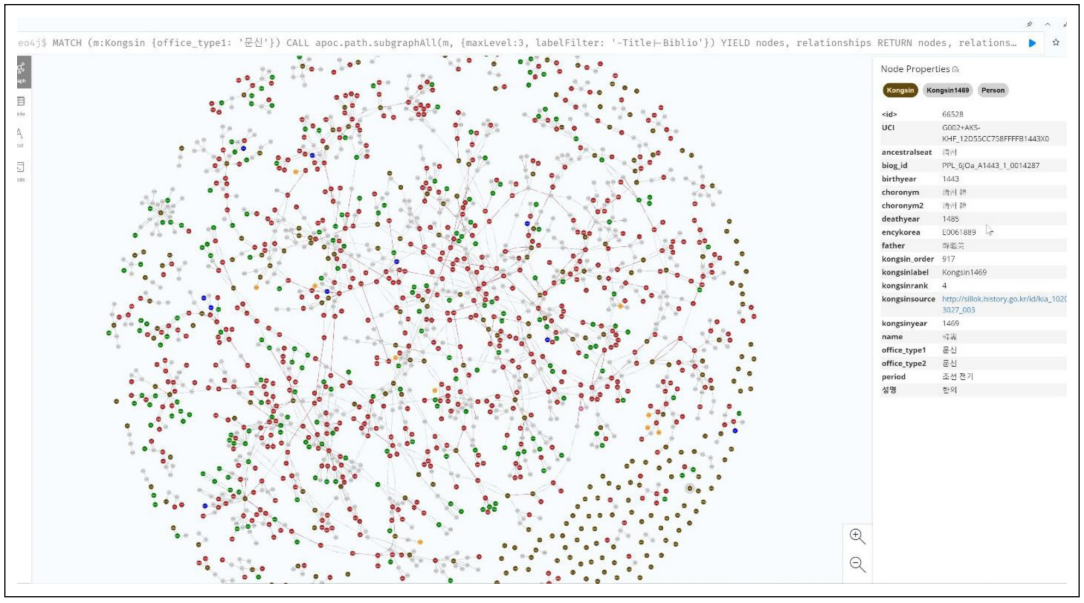


Figure 20: This figure shows a dense network of intergenerational and affinal connections among merit subjects of civil origins, suggesting the stable and gradual coalescence of some groups into self-perpetuating hereditary clusters.

Conclusion

Today, even the most traditional historians are more likely to begin their work in an online repository than in the library stacks or the proverbial dusty archive. The rise of searchable digital sources and bibliographic databases has reshaped the daily rhythms of historical practice. As Roy Rosenzweig observes, historians have shifted from grappling with scarcity to managing overwhelming abundance, all while dealing with the peculiar materiality of digital sources. This abundance brings unexpected possibilities: it allows us to move across archives and disciplines with new ease and, at times, to stumble onto connections through what Lara Putnam calls "side glancing" (380). In this sense, as Milligan aptly notes, we have all become digital historians ("We Are All Digital Now").

On the other hand, the digital turn also means historians face the challenge of developing sound methods for making sense of an ever-expanding body of digitized sources. Collections are more accessible than ever, but their sheer scale makes it difficult to move beyond surface-level discovery with conventional approaches. Keyword searches, though useful and convenient, tend to reinforce existing assumptions by returning only what is apparent and what we already expect to find. Topic modeling can provide a broad overview of a corpus, but it struggles to connect multiple, heterogeneous datasets and, because it relies on unsupervised learning, its results are

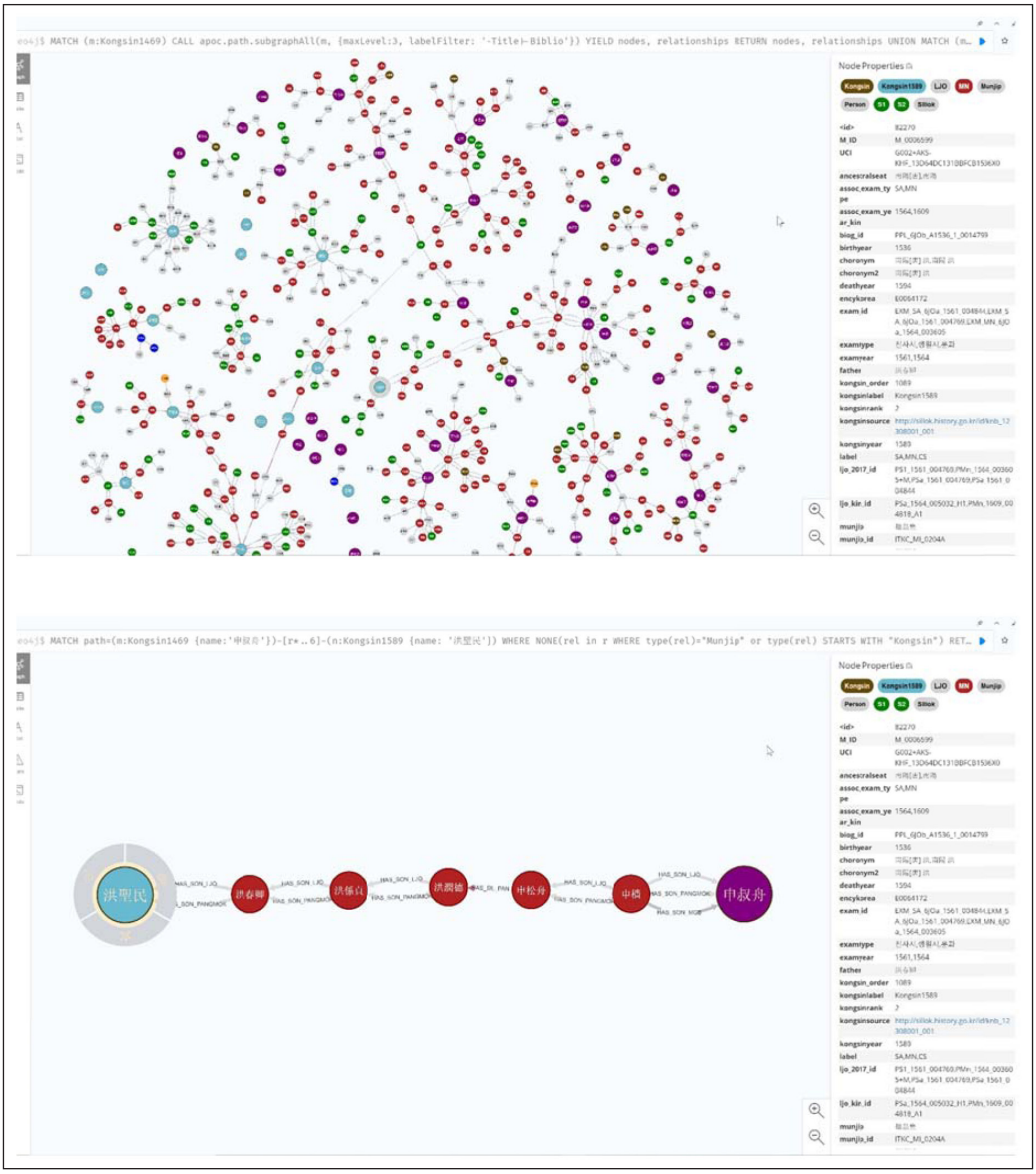


Figure 21: Multiple kinship chains span numerous nodes, including a long linkage between Sin Sukchu and Hong Sŏngmin. While such density does not necessarily indicate centrality, it highlights differences in evidentiary reliability across relationships.

not always consistent or reproducible—a major issue for the interpretive precision that historical research requires.

This article has argued that the next step for digital history should be to move toward machine-assisted inquiry that weaves together diverse sources in a critical and rigorous way. Visualization software such as Gephi and Cytoscape have been invaluable

for providing an intuitive aerial view, in a nod to Philip Bagby. Yet my early experiments with these tools revealed clear shortcomings in capturing the contextual specificity that Timothy Tangherlini has sought in computational folkloristics. Gephi's static graphs limit its effectiveness for tracing evolving aristocratic networks, whereas Cytoscape, though better at highlighting subnetworks and adjacent ties, reduces the integration of multiple datasets to a simple join on common key values. What historians need is a macroscope that balances breadth and depth, one that enables smooth movement between individual records, local clusters, and broader structural trends. For now, however, the off-the-shelf solutions available in digital history fall short of achieving that balance.

Historians do not simply gather information. We critically assess sources, reconcile conflicting accounts, and situate our findings in broader narratives and structures. We understand that centralized repositories and digital corpora are not permanent and authoritative stores of information, but rather imperfect collections with complicated backstories. A historian who aggregates data or visualizes relationships without embedding them in an interpretive framework risks becoming a digital antiquarian. Unlike computational literary studies, where a primary aim might be to examine the composition of a corpus itself, digital history works with extracted information that must be evaluated, corroborated, and contextualized in order to generate meaningful insights and arguments about past societies. In other words, in digital history, the corpus is a lens through which we develop our interpretations; it is not the object of research itself.

Looking ahead, several areas warrant further development. Large language models (LLMs) and vision-language models (VLMs) present both opportunities and challenges for historical research. Thus far, their most common applications have been in summarization through chatbot interfaces and automated text recognition. However, early experiments in my lab suggest that LLMs and VLMs could automate significant portions of a historian's typical workflow, particularly in the management of primary sources and data preprocessing. In preparation for a follow-up study to this article, I have successfully extracted all patronage and kinship information from medieval Korean epitaphs with remarkably high accuracy. In addition, recent experiments using VLMs and Chain-of-Thought (CoT) prompting indicate that we may not be far from replacing Putnam's "side glancing" with more sophisticated machine learning-assisted transnational and multilingual historical research.

Another important consideration is the role of open-source software and open data. Neo4j is available as a "community edition" under a modified GNU General

Public License, but it remains a commercial product. I adopted it as the backend of the Medieval Yangban Project primarily for reasons of speed, convenience, extensive documentation, and the active online communities that support it. That said, academic research should embrace openness wherever possible. In my defense, all of the data presented in this article are available as tab-separated values, which can be imported into any macroscope designed as a modular system. Looking ahead, I plan to implement a fully open-source solution, replacing closed-source Neo4j with ArangoDB or an equivalent database system.[15]

[15] I would like to thank one of the anonymous reviewers for pointing this out and for suggesting ArangoDB as an open-source alternative.

Acknowledgements

The data preprocessing and preparation for this article were generously supported by the Seoul National University Faculty Research Start-up Grant (990-20170013) and the 2018 and 2019 Data Science Advancement Initiative at Seoul National University's Big Data Institute (0660-20190012). My initial success in applying Neo4j to Korean historical research was made possible by my sabbatical stay as a Digital Historian-in-Residence at Lingnan University during the 2020–2021 academic year. The refinement of the research methodology benefited from the Hong Kong Research Grants Council's General Research Fund (17619523). I would like to thank Sang Kyun Cha and Vincent S. Leung for their institutional support and for placing trust in a highly experimental project that could have easily resulted in failure. Over the years, I have also received valuable input and suggestions from Eric Chow, Yan Hon Michael Chung, Christina Han, Jing Hu, Ian M. Miller, and Paul Vierthaler. Any remaining errors are my own.

Competing Interests

The author has no competing interests to declare.

Works Cited

Abello, James, Peter M. Broadwell, Timothy R. Tangherlini, and Haoyang Zhang. "Disentangling the Folklore Hairball: A Network Approach to the Characterization of the Large Folktale Corpus." *Fabula*, vol. 64, nos. 1–2, 2023, pp. 64–91, https://doi.org/10.1515/fabula-2023-0004.

Academy of Korean Studies. "Han'guk yŏktae inmul chonghap sisŭt'em." http://people.aks.ac.kr/front/board/info/introduction.aks. Accessed 15 February 2025.

Academy of Korean Studies. "Yŏktae inmul UCI." http://people.aks.ac.kr/front/uci/uciInfo.aks. Accessed 15 February 2025.

Bagby, Philip. *Culture and History: Prolomegna to the Comparative Study of Civilizations*. University of California Press, 1963 [1958].

Börner, Katy. "Plug-and-Play Macroscopes." *Communications of the ACM*, vol. 54, no. 3, 2011, pp. 60–69, https://doi.org/10.1145/1897852.1897871.

Cha, Javier. "Digital Korean Studies: Recent Advances and New Frontiers." *Digital Library Perspectives*, vol. 34, no. 3, 2018, pp. 227–44, https://doi.org/10.1108/dlp-04-2018-0013.

---. "Digital/Humanities: New Media and Old Ways in South Korea." *Asiascape: Digital Asia*, vol. 2, nos. 1–2, 2015, pp. 126–47, http://doi.org/10.1163/22142312-12340022.

---. "To Build a Centralizing Regime: Yangban Aristocracy and Medieval Patrimonialism." *Seoul Journal of Korean Studies*, vol. 32, no. 1, 2019, pp. 35–80.

China Biographical Database. "Rules for Index Years." https://cbdb.hsites.harvard.edu/file_url/700. Accessed 1 February 2025.

Crymble, Adam. *Technology and the Historian: Transformations in the Digital Age*. University of Illinois Press, 2021.

Graham, Shawn, Ian Milligan, and Scott Weingart. *Exploring Big Historical Data: The Historian's Macroscope*. Imperial College Press, 2015.

Hŏ Su. "*Kaebyŏk* nonjo ŭi sahoejuŭihwa e kwanhan saeroun chŏpkŭn? T'op'ik yŏn'gyŏlmang punsŏk ŭl chungsim ŭro [A New Approach to the Spread of Socialist Influence in *Kaebyŏk*? An Analysis of Topic Networks]." *Inmun nonch'ong*, vol. 78, no. 1, 2021, pp. 221–62.

Kim Hyeon [Kim Hyŏn]. *Inmun chŏngbohak ŭi mosaek* [In Search of Humanities Informatics]. Puk k'oria, 2012.

Lee, Sangkuk and Jong Hee Park. "Quality over Quantity: A Lineage-Survival Strategy of Elite Families in Premodern Korea." *Social Science History*, vol. 43, no. 1, 2019, pp. 31–61, https://doi.org/10.1017/ssh.2018.38.

Lee, Sangkuk and Wonjae Lee. "Strategizing Marriage: A Genealogical Analysis of Korean Marriage Networks." *Journal of Interdisciplinary History*, vol. 48, no. 1, 2017, pp. 1–19, https://doi.org/10.1162/JINH_a_01086.

Miller, Ian Matthew. "Rebellion, Crime and Violence in Qing China, 1722–1911: A Topic Modeling Approach." *Poetics*, vol. 41, no. 6, 2013, pp. 626–49, https://doi.org/10.1016/j.poetic.2013.06.005.

Milligan, Ian. "We Are All Digital Now: Digital Photography and the Reshaping of Historical Practice." *Canadian Historical Review*, vol. 101, no. 4, 2020, pp. 602–21, https://doi.org/10.3138/chr-2020-0023.

---. *The Transformation of Historical Research in the Digital Age*. Cambridge UP, 2022.

Park Hyun Soon [Pak Hyŏnsun]. "Kukcho munkwa pangmok ŭi p'yŏnch'an kwa 18 segi 'inmul chŏngbohak' [On the Publication of the Civil Service Examination Rosters and Eighteenth-Century 'Biographical Informatics']." *Kyujanggak*, vol. 56, 2020, pp. 173–215.

Putnam, Lara. "The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast." *American Historical Review*, vol. 121, no. 2, 2016, pp. 377–402, https://doi.org/10.1093/ahr/121.2.377.

Rosenzweig, Roy. "Scarcity of Abundance? Preserving the Past in a Digital Era." *American Historical Review*, vol. 108, no. 3, 2003, pp. 736–62, https://doi.org/10.1086/529596.

Tangherlini, Timothy R. "The Folklore Macroscope: Challenges for a Computational Folkloristics." *Western Folklore*, vol. 72, no. 1, 2013, pp. 7–27, https://www.jstor.org/stable/24550905.

Wagner, Edward W. "A Computer Study of Yi Dynasty Civil Examination Rosters," 22nd Annual Meeting of the Association for Asian Studies, 3 April 1970, San Francisco, CA.

---. *Literati Purges: Political Conflict in Early Yi Korea*. Harvard UP, 1974.

---. "Project Description." Box 30, Folder 3. Wagner–Song Correspondence, [1967–1997], Edward W. Wagner personal archive, Harvard University Archives.

---. "The Civil Examination Process as Social Leaven: The Case of the Northern Provinces in the Yi Dynasty." *Korea Journal*, vol. 17, no. 1, 1977, pp. 22–27.

---. "The Korean Chokpo as a Historical Source." In *Studies in Asian Genealogy*, edited by Spencer J. Palmer, Brigham Young UP, 1969, 141–52.