Digital Korean studies: recent advances and new frontiers

Javier Cha
Seoul National University, Seoul, South Korea

Received 1 May 2018
Revised 5 July 2018
Accepted 5 July 2018

Abstract

Purpose – This study aims to reflect on the past and prospects of digital Korean studies.

Design/methodology/approach – Discussion includes the remarkably early adoption of computing in the Korean humanities, the astounding pace in which Korean heritage materials have been digitized, and the challenges of balancing artisanal and laboratory approaches to digital research.

Findings – The main takeaway is to reconsider the widespread tendency in the digital humanities to privilege frequentist analysis and macro-level perspectives.

Practical implications – Cha hopes to discover the future of digital Korean studies in semantic networks, graph databases and anthropological inquiries.

Originality/value – Cha reconsiders existing tendencies in the digital humanities and looks to the future of digital Korean studies.

Keywords Digital Korean studies, Digital humanities, Korean studies, South Korea, Digital archives

Paper type Research paper

A century ago, the French historian Marc Bloch had a well-known exchange with the medievalist Henri Pirenne, at a conference held in Stockholm. As Bloch was about to head out to historical sites and museums, Pirenne set out to explore and soak in the modern and contemporary aspects of the Swedish capital. "If I were an antiquarian, I would have eyes only for old stuff", Bloch recalled Pirenne having said, "I am a historian, however, and therefore I take delight in the living" (Bloch, 1953/1941, pp. 43-47)[1]. The past, according to Pirenne, does not refer to the fossils of a bygone era but to that which is in a symbiotic relationship with the present day. That is, to the extent that the past enriches our understanding of the present, the past should be understood from the standpoint of the changing present situation. In a lecture I attended circa 2010, Carlo Ginzburg playfully remarked: "Anachronism is the most powerful tool available to the historian".

This article discusses how the anachronistic interplay between the past and the present might help create a new wave of scholarship in my primary area of research: the uses of

The open-access license fee has been paid for by a grant from the Office of Research Affairs at Seoul National University.

The author wishes to acknowledge the generous support provided by the Academy of Korean Studies Competitive Research Grant (AKS-2017-R05). Special thanks to Mark Byington, Kim Baro, Allan Cho, Kim Hyeon, John S. Lee, Rho Kyung Hee, Ryu Intae and Allison Van Deventer, and the anonymous *DLP* reviewer.

Digital Library Perspectives
Vol. 34 No. 3, 2018
pp. 227-244
Emerald Publishing Limited
2059-5816
DOI 10.1108/DLP-04-2018-0013

modern digital archives and computing power in the study of premodern Korea. In 2018, South Korea boasts large collections of heritage materials captured, archived and curated, using cutting-edge database technology. These databases have been made publicly downloadable under a government-mandated open license policy. This unusual situation, to my knowledge not found in any other area studies discipline, demands that Korea specialists think creatively and reflectively about the implications of having access to such a staggering amount of high-quality humanities data. How should these repositories be structured, curated, and preserved? In what ways do our existing interpretations of Korean history and culture change due to digitization and digital methods? What can digital Korean studies teach us about the advantages and limits of data-driven humanities? What would be some effective ways of incorporating images, audio recordings, aerial photography, and 3d scans of artifacts in the Korean humanities?

The field of Korean studies, especially outside of South Korea, has given surprisingly limited attention to computing and digital methods. This lack of interest stands in stark contrast to the South Korean government's massive investments in the building of archives and the attendant dependence of Koreanists on digitized source materials. At a recent Association of Asian Studies annual meeting, the inaugural digital humanities working group meeting was held; I found myself the only specialist of Korea in a room with approximately 60 attendees. I also turned out to be the only presenter covering Korea in the DH 2018 conference in Mexico City[2].

Nonetheless, there are some encouraging signs. Korean studies librarians have been paying close attention to South Korea's digitization efforts and the digital humanities[3]. In South Korean academia, the Korean Association for Digital Humanities (*Han'guk tijit'ŏl inmunhak hyŏbŭihoe* 한국 디지털 인문학 협의회) was established in 2015. In a few pockets, such as the Academy of Korean Studies, Hankuk University of Foreign Studies and Ajou University, the digital humanities have been gaining traction, albeit slowly (Kim, 2016, pp. 385-388). In 2016, a solid primer articulating a long-term vision, technical challenges, global comparisons, pedagogy and reflections on failed efforts was published (Kim *et al.*, 2016). To make sense of what "digital humanities" means in the South Korean context, however, it is important to keep in mind the legacy of a domestic phenomenon called "cultural contents studies" (*Munhwa k'ont'ench'ŭhak* 문화콘텐츠학). South Korea's impressive range of digital archives was funded primarily to foment the media industries, such as K-pop, television drama, cinema and video games, not necessarily to promote new types of research in the humanities.

The Korean case demonstrates that the availability of digitized materials, no matter how large in scale and how high in quality, does not spontaneously lead to explorations in new modalities enabled by digitization and digital technologies. Digital projects entail more than feeding big cultural data into a computer and expecting a groundbreaking result. Many, if not most, end up as failed experiments and lead researchers down unexpected and unforeseen paths. Months and years of training and tedious work are needed to produce meaningful outcomes and mature studies, which also necessitates cross-disciplinary cooperation with informatics, computer science, statistics and other relevant fields. In addition, digital humanities scholars need to be prepared to learn and embrace the aesthetic and storytelling aspects of digital media, such as photography, videography, graphic design, 3D modeling and animation and game engines.

Navigating the uncharted waters of digital Korean studies appears less daunting once we realize that the study of Korea's past has been influenced by modern digital technologies in more ways than generally recognized. Because of digitization, expectations have changed and continue to change. In 2006, for example, the release of *A Compendium of Korean Collected Works* (*Han'guk munjip ch'onggan* 韓國文集叢刊) as an online database was a groundbreaking moment for many historians of Korea. Only a portion of the 1,259 collected

works of premodern Korean intellectuals in the current database was made available then and the user interface was basic by today's standards[4]. Yet, I found this resource invaluable. I gained access to an expensive collection of primary sources that the University of British Columbia at the time could not afford to have in its library stacks due to budgetary and spatial constraints. Research that used to require three hours of driving from Vancouver to the University of Washington in Seattle or three weeks of waiting for one volume among hundreds to arrive on interlibrary loan could be done anywhere, as long as I had access to a computer connected to the internet. Accordingly, I was no longer satisfied with reading this collection one volume at a time, and I read the sources more extensively than before. As I got used to this online database's various keyword search functions, I experimented with a new approach that would have been difficult to execute without digitization: broad, comparative analysis of passages associated with a relatively obscure concept, covering the works of about two dozen scholars.

Fast forward to 2018: I now have this entire database on my notebook computer's solid-state drive as a 500-megabyte Unicode text file. The 18,398 raw XML files that make up this database can be downloaded legally and free of charge, under the terms of South Korea's Open License for public data, on the National Information Society Agency's Open Data Portal. With access to the entire database, my expectations are changing yet again. I look at the 557,126 pieces of writing consisting of 154 million characters in the 2017 version of this database and wonder what new type of research may be possible. Just to start, I have tried to run topic models, map out semantic patterns and algorithmically classify the authors and writings on the basis of diction, style and figures of speech.

To make sense of where digital Korean studies originated, where it stands today and where it is headed, we should recognize the foresight of predecessors who were ahead of their time. Computing in Korean studies had a remarkably early start owing to the pioneering efforts of Song June-ho (MR: Song Chunho, 1922-2003)[5] and Edward Wagner (1924-2001). In 1959, Wagner was appointed as an assistant professor of Korean history at Harvard University, upon finishing his doctoral dissertation on the political history of fifteenth- and sixteenth-century Korea at the same institution. He spent the initial years of his appointment laying a foundation of research and teaching about Korea in the United States, starting with the publication of a textbook in written Korean language in 1963, that would grow into a three-volume project over the years (Wagner, 1963/1971). His next project was to be a multi-volume introductory history of Korea composed of contributions from experts based in North America and Europe. That project never materialized, however, and he decided to pursue something else instead. In 1964, Song June-ho of Chonbuk National University, also a historian of Korea, visited the Harvard-Yenching Institute (Plate 1). During his year-long stay, Song persuaded Wagner to analyze the elite structure of the Chosŏn 朝鮮 dynasty (1392-1910) using the roster of civil service examination degrees, called *munkwa* 文科. At the time, Song and Wagner did not realize that what later came to be known as the Munkwa Project would involve computing, nor that their collaboration would last for nearly 40 years. In 1966, Song began collecting portions of the examination roster in Japan[6], and a successful Ford Foundation grant application in 1967 officially initiated the Munkwa Project with the promise of compiling a database. This development in Korean studies took place during what the French historian Emmanuel Le Roy Ladurie called the "American challenge" in reference to the vigorous push in the United States to install computers on university campuses (Ladurie, 1979/1968, p. 6).

The Munkwa Project inspired the creation of other digital archives, but the resources required for such projects became abundant as the Government of South Korea's response to the 1997 Asian financial crisis included digitization. In 1998, an economic stimulus program called the Informatization Labor Project (*Chŏngbohwa kŭllo saŏp* 정보화근로사업) initially

Plate 1.
Edward Wagner and Song June-ho at a Buddhist temple near Chŏnju in 1970, along with their respective family members (from the Edward W. Wagner personal archive at Harvard University Archives)

allocated approximately USD $200m over two years to create 48,000 white-collar jobs involving the digitization of cultural heritage (Kim, 2012, p. 601; Cha, 2015, p. 139). The National DB super-collection, an outgrowth of that initiative, now lists thirty-one databases categorized under "history" (*yŏksa* 역사), another set of thirty-one under "culture" (*munhwa* 문화), and 16 under "education" (*kyoyuk haksul* 교육학술)[7]. In addition to these 78 obvious candidates, several others filed under "science and technology" (*kwahak kisul* 과학기술) and "industry and economy" (*sanŏp kyŏngje* 산업경제), such as climate, public health, science and technology magazines, satellite images and bags of words of Korean and major world languages, should be of interest to humanists and social scientists as well. Nearly USD $1 billion has been collectively spent on National DBs since its inception as the Informatization Labor Project twenty years ago. Large amounts of public funds are continuing to drive the building of big databases. In 2017, the National Institute of Korean Language was granted USD $17.5 million over five years to create 15.5 million bags of words representing the modern Korean language for AI-driven linguistic analysis[8]. At the Institute for the Translation of Korean Classics, USD $20 million is being invested annually to train a deep-learning model for translating *The Diary of the Royal Secretariat* (Sŭngjŏngwŏn ilgi 承政院日記). Prior to this, the digitization of the scribes' notes on the daily affairs of the early modern Korean court, covering the years from 1623 to 1910 in 242 million Sinitic characters, took 15 years, from 2001 to 2015[9]. This painstaking task required the deciphering of documents written in cursive and shorthand forms. The next step of rendering the literary Chinese content into modern Korean is projected to take at least 45 years with the compensation for the necessary specialists limited to a meager USD $15 per page due to the project's scale[10]. Using the new deep-learning approach, the estimated time for completing the translation has been reduced to 18 years, at an annual cost of what one journalist

jokingly remarked was the equivalent of "the asking price of a single apartment unit in Gangnam [MR: Kangnam][11]".

In addition to benefiting from an early start and large-scale funding, South Korea's machine-readable archives tend to be of remarkably high quality. For this, digital Koreanists are indebted to one of the trailblazers: Kim Hyeon (MR: Kim Hyŏn), who currently heads the department of humanities informatics at the Academy of Korean Studies. Originally a specialist of Korea's Neo-Confucian philosophy, Kim Hyeon's foray into humanities computing began in 1985 with a position in the Korea Institute of Science and Technology. His initial interest in informatics involved the encoding of *han'gul* 한글 (modern Korean phonetic characters) and *hancha* 漢字 (a set of Sinitic characters used in written Korean language). Throughout his distinguished career in humanities computing and digital humanities, he had the extraordinary ability to remain at once obstinate and free from dogmatism about technology. In the early 1990s, when CD-ROM emerged as the high-capacity storage medium of the future, he helped found a start-up company to produce the first-ever digital edition of the *Annals of the Chosŏn Dynasty* (*Chosŏn wangjo sillok* 朝鮮王朝實錄), offering the ability to search through its 50 million characters of full text. During the years of South Korea's rapid expansion of digital infrastructure, his company transferred the ownership rights of this database to the National Institute of Korean History. Throughout this process and in a different capacity, Kim played a key role in reworking the archive's data ontology for the internet. The current online edition of the *Annals of the Chosŏn Dynasty* consists of 674 XML documents, with detailed annotations of every full-text entry (Figure 1). This innovative design allows the researcher, for example, to query the information regarding 336,267 official career records mentioned in the court records with the ability to provide citations for each and every entry (Figures 2 and 3). A simplification of this XML schema was ported to the aforementioned *A Compendium of Korean Collected Works* and used to structure the writings of 1,259 authors (Figure 4).

Moreover, digital archives in Korean studies are remarkably up-to-date in database and content-management technologies. Beyond XML, Kim Hyeon has avidly adopted, promoted and experimented with the wiki platform, the semantic web, and Neo4j's GraphDB, among others. Many recent academic projects related to Korean studies and digital humanities at the Academy of Korean Studies, such as the annual training workshop in reading and translating literary Chinese sources and a recent conference on digital storytelling, are

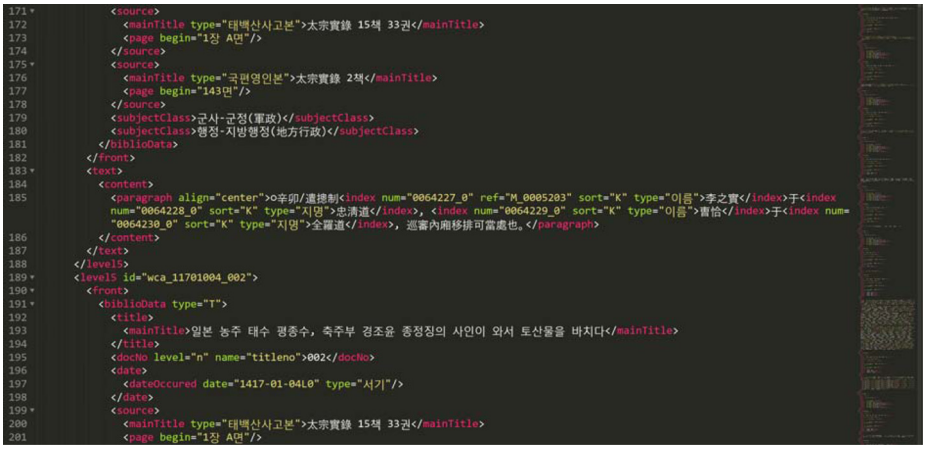


Figure 1.
A heavily XML-tagged Sillok entry

인물정보 상세보기

인물정보 : "서거정(徐居正)"

실록에서 검색하기

기본정보

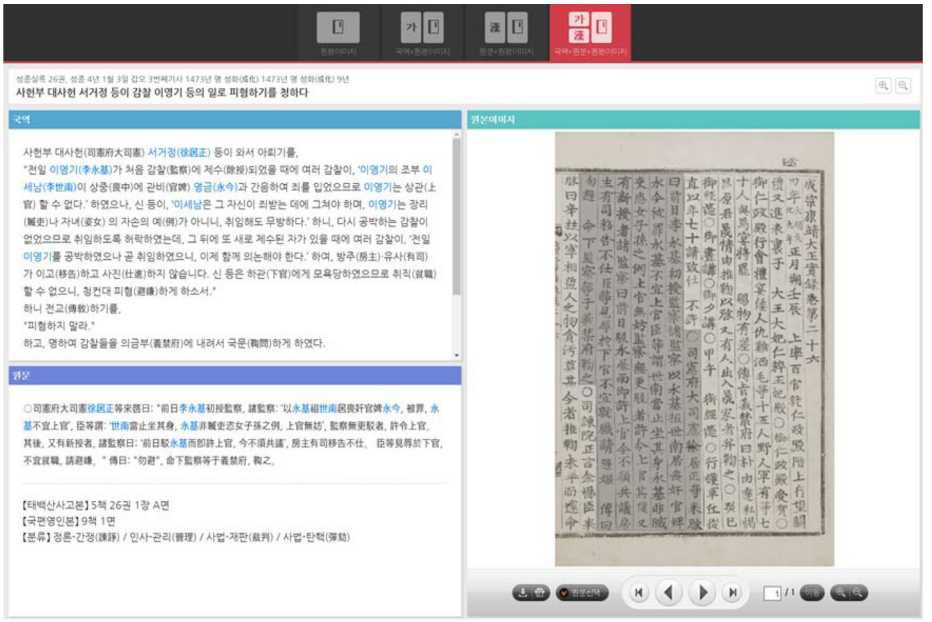
이름	서거정(徐居正)
본관	대구(大丘)
성씨	서(徐)
자	강중(剛仲)
호	사가정(四佳亭)
시호	문충(文忠)
생년	1420
몰년	1488
신분	문반

관련

왕대	왕력	년	월	일	관력	형태	해당기사
세종	26	1444	10	11	박사(博士)	현직	바로가기
세종	28	1446	10	10	박사(博士)	현직	바로가기
문종	0	1450	9	7	수찬(修撰)	현직	바로가기
문종	0	1450	9	27	수찬(修撰)	현직	바로가기
문종	1	1451	5	25	사경(司經)	현직	바로가기
문종	1	1451	11	25	집현전 부교리(集賢殿副校理)	현직	바로가기
세조	1	1455	9	2	응교(應敎)	현직	바로가기
세조	1	1455	12	27	응교(應敎)	현직	바로가기
세조	2	1456	9	11	예문 응교(藝文應敎)	현직	바로가기
세조	3	1457	2	9	우사간 대부(右司諫大夫)	현직	바로가기

Figure 2.
The career history of Sŏ Kŏjŏng (1420-1488) generated via real-time query request on the XML files that make up *The Annals of the Chosŏn Dynasty*

published as collaborative wiki content[12]. Such contents, as well as the various dictionaries, concordances and other reference materials converted into digital form, have their semantic properties and relational features tagged. Network visualizations are embedded into the wiki publications themselves to enable discovery and navigation into pertinent contents. As for GraphDB, Kim Hyeon and his students and associates in humanities informatics have demonstrated the benefits of utilizing such an intuitive, robust and flexible database technology to digitize humanities data. To mention one recent



Note: One click brings the user to the court entry linked to that official career information, which shows "Chief Censor Sŏ Kŏjŏng" in modern Korean translation, digitized Sinitic text, and scanned image of the page in which this court entry appears

Figure 3.
On the third day of the first lunar month of 1473, Sŏ Kŏjŏng was the Chief Censor

successful example, the Academy of Korean Studies has converted the genealogy of the Chosŏn dynasty's royal family members and their extended kin, called Sŏnwŏllok 璿源 into GraphDB. The data set consists of approximately 596,000 nodes and 767,000 edges, waiting to be used by researchers (Figure 5).

Digital Korean studies is entering the realm of big data. To handle digitized archives of growing size and complexity, Korean studies will need to consider transitioning from an artisanal to a laboratory mindset[13]. Yet, Koreanists still overwhelmingly prefer the artisanal approach of producing specialized and detailed case studies of individuals, local communities and institutions by excavating previously unstudied or understudied materials. The belief is that the accumulation of research conducted in this manner eventually will result in a new understanding of Korea's past, one that is more objective than previous interpretations and more faithful to what the primary sources show. Collectively speaking, artisanal Korean studies is grounded in the assumption that case studies follow a normal distribution of the individuals, topics and local communities that make up the Korean peninsula and the Korean diaspora. Some influential figures, famous communities and important topics may receive more attention than others, but overall such unevenness can be corrected.

Today's age of big cultural data, however, turns this assumption on its head. The actual picture is far more skewed than what we might imagine. Along with the digitization of primary sources, secondary studies have been made available in digital form via three major service providers: KISS, DBpia and RISS. KISS has built a database with 1,387,413 articles since 1996[14]; DBpia has accumulated 2,221,278 articles, 19,630 e-books and 31,916

Figure 4.
A portion of the digital edition of *Sŏngsobu pugo*, or the collected writings of Hŏ Kyun (1569-1618)

Note: The XML annotation is a simplification of the schema originally utilized in *The Annals of the Chosŏn Dynasty*. The tagging is sparser compared to the example shown on Figure 1. There are plans to markup more details in each document

references since 2000[15]; and RISS has 4,807,098 articles and 1,382,304 dissertations since 1996[16]. It has become possible to survey the distribution of case studies with the aid of macro software (Figure 6). For example, *A Compendium of Korean Collected Works*, a database mentioned earlier, contains the writings of 1,259 premodern Korean authors. How

many of those authors have been studied? In a survey of 530 authors born between 1450 and 1750, I sought to estimate the number of case studies published on each author available on KISS. The query returned approximately 2,476 entries, with a negligible number of entries erroneously counted and missing. The visualization of this data set rendered a Pareto chart that demonstrates the extreme inequality of this power-law distribution (Figure 7). Yi Hwang 李滉 (1501-1570), perhaps the most famous Korean Neo-Confucian scholar, leads the pack with 578 case studies or 23 per cent of the total. Next in line is his rival in the south, Cho Sik 曺植 (1501-1572), with 116 case studies or 4.7 per cent (Table I). Overall, 16 authors made up half of all case studies available on KISS, and the 80 per cent mark was reached with only 77 authors or 14.5 per cent of 531 authors. The most disconcerting finding of this exercise is the long trail of zeroes: 248 authors, or 46.7 per cent, have yet to be the subject of

Figure 5. The Chosŏn dynasty's royal family genealogy in GraphDB, which consists of approximately 596,000 nodes and 767,000 edges

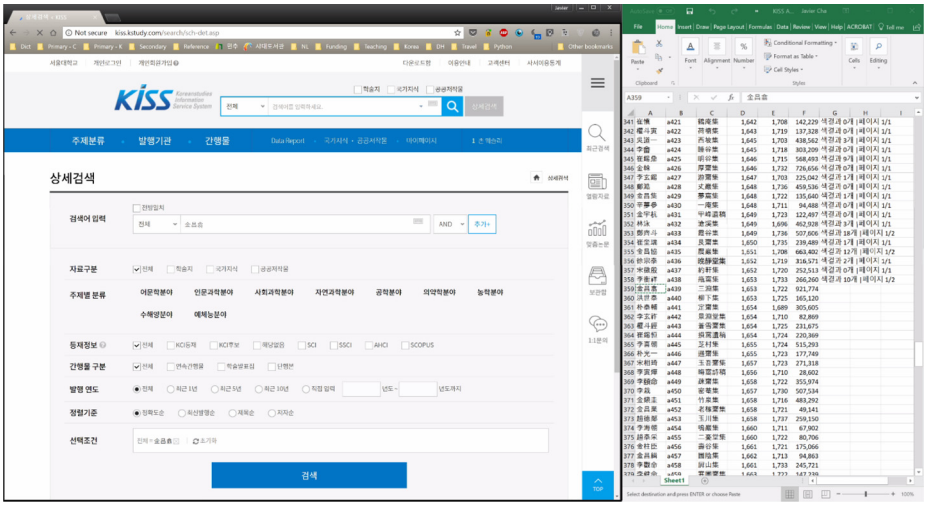


Figure 6. Scraping bibliometric data by batch-running advanced search queries on KISS using Macro Express Pro

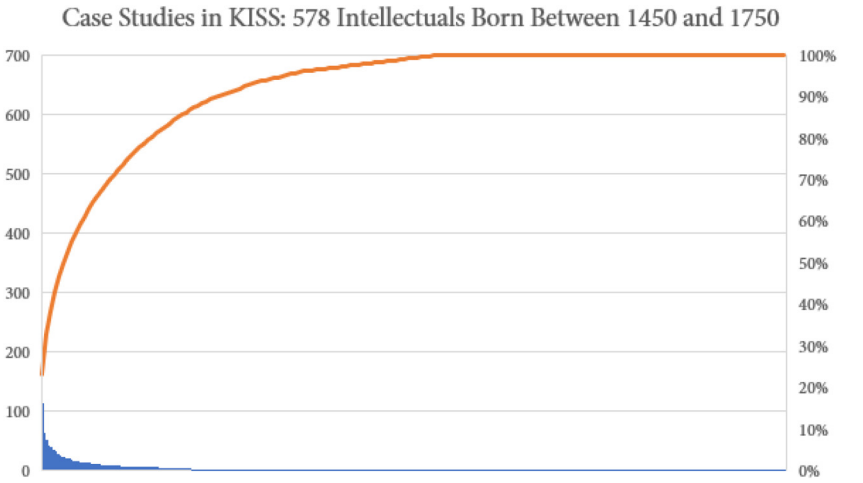


Figure 7.
A Pareto chart of case studies that appear in the KISS database for Korean intellectuals born between 1450 and 1750

case study, according to the KISS records. Put in another perspective, 405 authors, or 76.3 per cent, have three or fewer case studies written about them. Numbers do not tell the whole story, of course, but these figures are certainly eye-opening.

Intervention is necessary. The artisanal approach has evidently resulted in a highly uneven picture of Korea's past. What should be done? What needs to be done? My initial response was to experiment with macro-level text analysis. Along with a quantitative sociologist, I tried to run topic models and identify latent patterns in the data. This turned out to be more problematic than we had anticipated. Unlike the relatively consistent corpus that made possible Jockers (2013) "macroanalysis" of British, Irish and Irish American literature, Korean collected works consist of a mix of various prose and verse forms that appear deceptively to be uniform due to our prejudices about the timelessness of Sinitic writing[17]. In addition, we ran into segmentation issues: poetry in Chinese and Korean could not be broken down into meaningful units and prose in literary Chinese had no parser readily available. Most of the results of our topic modeling attempts returned gibberish. Segmenting the writing into 2-grams, 3-grams and 4-grams using dictionaries helped somewhat but did not address the fundamental challenges of working with the peculiar features of our data set (Figure 8).

As an alternative strategy, I attempted what Jockers did with function words on English-language corpora (Jockers, 2013, p. 65), but on a small subset of literary Chinese prose written in the neoclassical mode, called Tang-Song "ancient prose" style (Korean *komun*/Chinese *guwen* 古文) in East Asian literature. With the input of a Korean literature specialist, I loaded the prose writings of four early seventeenth-century literary masters: Yi Chŏnggwi 李廷龜 (1564-1635), Sin Hŭm 申欽 (1566-1628), Chang Yu 張維 (1588-1638) and Yi Sik 李植 (1584-1647). The four masters, known as Wŏl Sang Kye T'aek 月象谿澤 by the initials of their pen names, are renowned for their elegant prose in the Tang-Song neoclassical mode. However, Yi Chŏnggwi and Yi Sik have been suspected of being influenced by a contemporary literary trend that became fashionable in Beijing: archaism (Korean *pokko*/Chinese *fugu* 復古) or Old Phraseology (Korean *komunsa*/Chinese *guwenci* 古文辭) (Bryant, 2008; Chang, 2010, pp. 28-36; Rho, 2015; Ong, 2016). Simply put, archaists sought to make neoclassical prose "truly" ancient and one of the ways to do that was to make prose more "lyrical" by suppressing the use of function words and grammatical

		(%)
李滉	578	23.34
曺植	116	28.03
朴趾源	64	30.61
李珥	54	32.79
宋時烈	52	34.89
李德懋	43	36.63
李瀷	41	38.29
洪大容	41	39.94
安鼎福	37	41.44
許筠	36	42.89
朴世堂	35	44.31
尹拯	29	45.48
黃胤錫	29	46.65
李恒	27	47.74
申欽	25	48.75
鄭逑	24	49.72
朴齊家	24	50.69
韓元震	23	51.62
張顯光	22	52.50
許穆	22	53.39
趙憲	21	54.24
丁時翰	19	55.01
鄭仁弘	18	55.74
張維	18	56.46
鄭齊斗	18	57.19
趙靖	17	57.88
趙翼	17	58.56
宋純	16	59.21
梁應鼎	16	59.85
李玄逸	16	60.50
趙光祖	15	61.11
奇大升	15	61.71
尹善道	15	62.32
宋浚吉	15	62.92
安錫儆	14	63.49
朴世采	13	64.01
徐命膺	13	64.54
宋翼弼	12	65.02
姜沆	12	65.51
金昌協	12	65.99
李柬	12	66.48
申維翰	12	66.96
申光漢	11	67.41
成渾	11	67.85
鄭澈	11	68.30
李達	11	68.74
李恒福	11	69.18
趙纘韓	11	69.63
李植	11	70.07
南孝溫	10	70.48
曺偉	10	70.88
李彦迪	10	71.28

(*continued*)

Table I.
The data table that make up the pareto chart on Figure 7

		(%)
盧守愼	10	71.69
朴淳	10	72.09
趙絅	10	72.50
李衡祥	10	72.90
林億齡	9	73.26
林薰	9	73.63
李楨	9	73.99
曺好益	9	74.35
郭再祐	9	74.72
崔錫鼎	9	75.08
李夏坤	9	75.44
李匡師	9	75.81
申景濬	9	76.17
蔡濟恭	9	76.53
鄭汝昌	8	76.86
高敬命	8	77.18
李廷龜	8	77.50
李用休	8	77.83
魏伯珪	8	78.15
崔溥	7	78.43
金馹孫	7	78.72
朴祥	7	79.00
徐敬德	7	79.28
周世鵬	7	79.56
林芸	7	79.85
柳夢寅	7	80.13
鄭經世	7	80.41
鄭蘊	7	80.69
金尙憲	7	80.98
李安訥	7	81.26
申翊聖	7	81.54
申光洙	7	81.83
柳希春	6	82.07
文益成	6	82.31
白光勳	6	82.55
李德弘	6	82.79
柳成龍	6	83.04
金長生	6	83.28
金堉	6	83.52
崔鳴吉	6	83.76
李惟泰	6	84.01
南龍翼	6	84.25
南九萬	6	84.49
李象靖	6	84.73
任聖周	6	84.98
李忠翊	6	85.22
柳得恭	6	85.46
李浚慶	5	85.66
趙穆	5	85.86
鄭琢	5	86.07
權好文	5	86.27

Table I.

(*continued*)

		(%)
崔慶昌	5	86.47
吳瑗	5	86.67
崔興遠	5	86.87
李麟祥	5	87.08
李穆	4	87.24
黃俊良	4	87.40
權擘	4	87.56
朴光前	4	87.72
趙宗道	4	87.88
李元翼	4	88.05
高尙顔	4	88.21
李敏求	4	88.37
李惟樟	4	88.53
洪世泰	4	88.69
李光庭	4	88.85
南克寬	4	89.01
吳光運	4	89.18
趙龜命	4	89.34
兪漢雋	4	89.50
李家煥	4	89.66
金載瓚	4	89.82
朴雲	4	89.98
30 authors: 3 studies or fewer		
31 authors: 2 studies or fewer		
96 authors: 1 study		
248 authors: 0 study		

Table I.

particles. Thus, an unusually low occurrence of some of the most common function words and grammatical particles in literary Chinese prose could be interpreted as a sign that the author might have been under the influence of sixteenth-century archaism. While the method should be refined, a preliminary analysis of Wŏl Sang Kye T'aek prose pieces on MARKUS has revealed that Yi Chŏnggwi and Yi Sik indeed show a tendency to suppress the use of common function words and grammatical particles, at 9.9 and 9.7 per cent of all characters, respectively, compared to 12 to 17.7 per cent shown in the writings of other authors not suspected to have been archaists (Table II). Parenthetically, I attempted this exercise on Voyant but realized that the algorithm segmented the text into characters and words using, what I suspect was, a modern Chinese language parser. I ended up with erroneous results and there was no way to switch off the automatic parsing.

Frequentist analysis of humanities data can be useful. However, its limits should be acknowledged as well. Perhaps the strategies for identifying authors by their habits and influences can be scaled by using existing software tools and creating algorithms that automate the analyses and comparisons. The trouble is that the database itself is skewed. Those among the 1,259 authors whose case studies are overrepresented happen to have a large number of their writings preserved. On a timeline, the overrepresentation is concentrated in the seventeenth and eighteenth centuries. In *A Compendium of Korean Collected Works*, the total character length of writings before 1375, the bin range that roughly corresponds to the dynastic change from Koryŏ (918-1392) to Chosŏn, is 758,687, or only 5 per cent of the 150 million characters that constitute this database. The application of methods such as principal

```
4$0$lda$fv1$4$0.011*"雖無" + 0.010*"不是" + 0.007*"痛哭" + 0.007*"寂寞" + 0.007*"無人"
5$0$lda$fv1$5$0.011*"爲有" + 0.009*"近來" + 0.007*"萬里" + 0.007*"那堪" + 0.007*"悵望"
6$0$lda$fv1$6$0.009*"一聲" + 0.008*"力於" + 0.008*"處分" + 0.008*"心於" + 0.008*"故云"
7$0$lda$fv1$7$0.007*"而有" + 0.005*"一見" + 0.004*"上之" + 0.004*"所未" + 0.004*"爲先"
8$0$lda$fv1$8$0.011*"十分" + 0.007*"悠悠" + 0.006*"惆悵" + 0.006*"有時" + 0.006*"周公"
9$0$lda$fv1$9$0.012*"明日" + 0.012*"一笑" + 0.011*"無人" + 0.009*"悠悠" + 0.009*"何妨"
10$0$lda$fv1$10$0.008*"高樓" + 0.007*"尙有" + 0.007*"伏惟" + 0.007*"之言" + 0.006*"誰知"
11$0$lda$fv1$11$0.013*"主人" + 0.012*"此行" + 0.010*"春來" + 0.009*"與公" + 0.007*"天意"
12$0$lda$fv1$12$0.016*"不必" + 0.013*"月明" + 0.009*"淸風" + 0.006*"去不" + 0.005*"世路"
13$0$lda$fv1$13$0.010*"國者" + 0.010*"山下" + 0.010*"三月" + 0.009*"異日" + 0.008*"使君"
14$0$lda$fv1$14$0.006*"故人" + 0.006*"九重" + 0.005*"不可" + 0.005*"看書" + 0.005*"之禮"
15$0$lda$fv1$15$0.008*"先生" + 0.007*"一番" + 0.006*"作一" + 0.006*"無狀" + 0.006*"如何"
16$0$lda$fv1$16$0.014*"一聲" + 0.007*"得爲" + 0.007*"偶然" + 0.007*"出來" + 0.007*"不禁"
17$0$lda$fv1$17$0.011*"一時" + 0.008*"花開" + 0.008*"安之" + 0.008*"能無" + 0.006*"平生"
18$0$lda$fv1$18$0.018*"山色" + 0.012*"靑山" + 0.010*"此時" + 0.008*"胷中" + 0.008*"今朝"
19$0$lda$fv1$19$0.012*"楊州" + 0.012*"人家" + 0.010*"朝家" + 0.009*"聞公" + 0.009*"何處"
20$0$lda$fv1$20$0.011*"先生" + 0.009*"不必" + 0.009*"蒼茫" + 0.007*"秋風" + 0.007*"蕭然"
21$0$lda$fv1$21$0.008*"在天" + 0.008*"奈何" + 0.007*"春色" + 0.006*"無人" + 0.006*"黃花"
22$0$lda$fv1$22$0.010*"光陰" + 0.009*"未爲" + 0.007*"江山" + 0.006*"未死" + 0.006*"言有"
23$0$lda$fv1$23$0.006*"公有" + 0.006*"有一" + 0.006*"從前" + 0.005*"不覺" + 0.005*"起居"
24$0$lda$fv1$24$0.011*"天寒" + 0.011*"切於" + 0.009*"何忍" + 0.008*"千里" + 0.007*"風流"
25$0$lda$fv1$25$0.009*"千里" + 0.009*"一箇" + 0.008*"望之" + 0.008*"造物" + 0.006*"江南"
26$0$lda$fv1$26$0.012*"四海" + 0.008*"應有" + 0.008*"歸田" + 0.008*"耿耿" + 0.007*"白雪"
27$0$lda$fv1$27$0.014*"何日" + 0.009*"光陰" + 0.007*"人得" + 0.007*"歸路" + 0.005*"一事"
28$0$lda$fv1$28$0.011*"相逢" + 0.009*"歸去" + 0.008*"珍重" + 0.006*"恐有" + 0.006*"落花"
29$0$lda$fv1$29$0.009*"亦何" + 0.008*"大而" + 0.007*"萬事" + 0.007*"終日" + 0.007*"而長"
```

Figure 8.
A failed attempt at running a topic model (LDA) on 150 million Sinitic characters

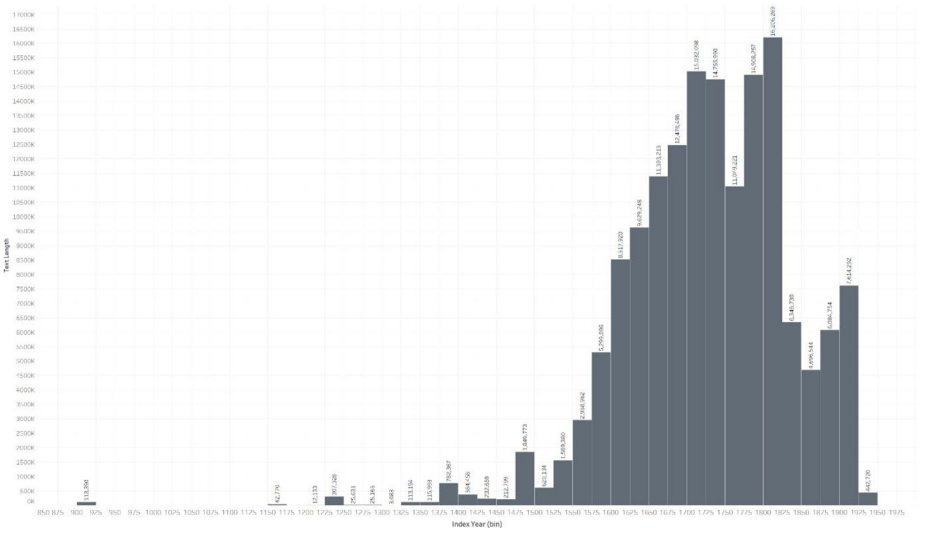


Figure 9.
The grossly uneven distribution of texts, by total character count, in *A Compendium of Korean Collected Works* database

component analysis and topic modeling, which require a sizeable corpus, makes sense for Chosŏn-era data but not Koryŏ. I am a specialist of Koryŏ. Nevertheless, my digital projects and experiments so far have almost exclusively involved Chosŏn data sets.

My mixed experience with frequentist approaches has led me to think about digital methods in terms of linkages and connections. Coming full circle, I have returned to what

Song June-ho and Edward Wagner intended to do with the Munkwa Project. I have also come to develop a genuine appreciation for the value of Kim Hyeon's preferred mode of developing digital Korean studies using wikis and semantic webs. The impressive digital infrastructure available to Koreanists drives the temptation to go for top–down, omniscient observations. However, I would argue that seeing "everything" is the easy part. Song and Wagner managed to computerize 14,600 records and aggregated the data by categories such as address, choronym, data and exam performance in only two or three years, without the power and ease of today's digital technology. The Munkwa Project ended up being an unfinished 40-year-old enterprise because seeing "everything" in this sense was not the goal. Song and Wagner sought to examine the rich contours of premodern Korea's elite structure by linking the exam degree roster with the vast pool of information stored in genealogies. Similarly, Kim Hyeon's vision is to organize the existing knowledge base of Korean studies in network representations and to create digital environments that encourage scholarly collaboration. As someone who has been involved in numerous digitization projects, Kim Hyeon had many opportunities to seek omniscience. Yet, he has shown little interest in such endeavors. Why?

Digital projects inspire researchers to try strange things. In 2011, Kim Hyeon obtained a pilot license (Kim, 2012). What motivated him to fly? In his words

> To create hypermedia contents that vividly capture Korea's local cultures, I decided to grab the control stick of a light aircraft. I did this for the same reasons I became a programmer and held a camera for the first time. At first, I couldn't help but laugh. Even I thought I was taking things too far [. . .]. However, as I pondered this issue for three or four months, my reasons for flying became clearer. (Kim, 2012, p. 828)

	Yi Chŏnggwi	Sin Hŭm	Chang Yu	Yi sik	Sŏ Kŏjŏng	Yi saek
Total characters	403,907	181,316	69,310	200,637	99,334	122,417
Average	588	553	587	822	689	532
Median	321	288	555	467	589	423
Pieces	687	328	118	244	145	230
Ratio (%)	9.92	12.01	17.66	9.69	12.43	13.98
Function characters	40,067	21,774	12,243	19,441	12,352	17,115
也	2,412	2,178	1,244	1,309	828	2,156
矣	1,341	710	417	606	390	882
之	13,176	5,942	3,118	5,630	4,349	4,862
者	2,338	2,053	1,132	1,484	1,141	1,019
而	5,071	3,108	1,823	2,458	1,215	1,617
以	5,452	2,315	1,226	2,914	1,137	1,660
所	2,114	857	561	1,149	469	934
其	3,319	2,042	1,152	2,037	1,083	2,030
於	3,767	1,791	888	1,338	1,202	1,057
乎	705	524	401	279	393	519
焉	372	254	281	237	145	379
也 (%)	0.60	1.20	1.79	0.65	0.83	1.76
也 + 而 (%)	1.85	2.92	4.43	1.88	2.06	3.08

Notes: Sin Hŭm and Chang Yu are their contemporaries. Sŏ Kŏjŏng 徐居正 (1420-1488) and Yi Saek 李穡 (1328-1396) are famous neoclassical prose writers whose works have been analyzed here for comparison

Table II. Evidence of suppressing function characters and grammatical particles in the prose writings of Yi Chŏnggwi and Yi Sik, who are suspected of having been influenced by the old phraseology movement. Sin Hŭm and Chang Yu are their contemporaries

Today, I use a sub-$1,000 drone to do a task that not long ago used to require a pilot license and an airplane. I fly a drone for the same reasons that brought Kim Hyeon to the sky on an airplane: exploration and discovery. The impressive digital infrastructure in Korean studies has made it possible to write an entire article without having to leave my desk. Yet, I have found myself, more than ever, actively engaging in field work, which is unusual for a specialist of the medieval and early modern periods. When I visit Andong, Chinju or Tamyang, I realize how little I know about the environment in which my historical subjects lived. I want to be in the same location where the local poets' societies gathered. I want to know the common routes that connected the various settlements in the area. I want to experience life in the region during different seasons. Most importantly, I want to see their world, despite the gap between my time and their time, from as many different angles as possible. This desire has propelled me to develop a serious interest in photography and videography and to think about framing and capturing the world around me in field sites from multiple angles and using lenses of various formats and apertures. I also learned how to fly a drone to access macro-level perspectives of a different nature from those I get from running topic models on a textual corpus. Every time I have tried something, I gained new insights.

Fortunately, the field of digital Korean studies seems to be headed in this direction. That is, one which prioritizes bridging the gap between the life and times of our historical subjects and we modern-day researchers equipped with digitized archives, cameras, sensors and computing power. In this emergent paradigm, our collective pursuit is not omniscience but immersion and connections. Recently, I was asked to join a team of senior academics, graduate students and database experts to create a database of citations, classical references and various instances of text reuse in the collected works of Koryŏ authors. My initial reaction to this project was skepticism: why not use text-reuse algorithms to detect such text-reuse patterns? Gradually, I was sold on the beauty of this project. The final project proposal ended up consisting of specialists with a wide range of expertise and interests but united under a common goal: to explore and discover new meanings and connections in the sparsely surviving Koryŏ-era writings, which, as aforementioned, make up only 5 per cent of the *A Compendium of Korean Collected Works* database. No text-reuse algorithm is substitute for a room of experts who can distinguish, for example, whether a classical reference was made directly or by way of other reference materials or the works of influential Chinese figures. As South Korea's full-fledged effort to digitize cultural heritage enters its twentieth year since the Informatization Labor Project during the Asian financial crisis, there have been concerns that we are "running out" of Koryŏ-era materials to digitize. The idea of building text-reuse databases shows an alternative path that could become a model for digital humanist scholars of other periods and other parts of the world. The next-generation databases for digital Koreanists will have the potential to showcase a new kind of concordance through which we can map the links and flows in the transformation of cultures over time. To do this, what digital Korean studies needs is not simply a shift from a field consisting of artisans to a field consisting of laboratories, but a field consisting of many laboratories of artisans. The field particularly needs eccentric artisans who might show up to field sites in airplanes.

Notes

1. I have introduced some minor changes to Peter Putnam's translation of this passage.
2. The DH 2018 conference program is available at: https://dh2018.adho.org/en/talleres/

3. In November 2017, the University of Michigan hosted the 2017 Workshop on Korean Data Services. Panel sessions included digital humanities, text mining, GIS and macro programming. In April 2018, Mikyung Kang, who oversees the Korea collection at Harvard-Yenching Library, and Yunah Sung, the Korean Studies Librarian at the University of Michigan, attended the DHAsia Summit at Stanford University. Nadia Kreeft, who curates the Korean materials at Leiden University Libraries, has been involved with the digital humanities initiatives at her home institution as well.
4. According to Kim H., 2012, p. 517, approximately 550 collected works were digitized by the end of 2005.
5. In this article, Korean names are given in their most common transliteration. The standard McCune-Reischauer romanization is provided following the label "MR".
6. Available at: http://nrs.harvard.edu/urn-3:HUL.ARCH:32836464?n=2
7. Available at: http://koreadb.data.go.kr/frt/ctl/sphereDB/selectSphereList.do?fieldCode=0
8. Available at: www.yonhapnews.co.kr/bulletin/2017/10/08/0200000000AKR20171008048600005.HTML
9. Available at: http://m.news.naver.com/read.nhn?mode=LSD&sid1=001&oid=001&aid=0007980956

   http://m.news.naver.com/read.nhn?mode=LSD&sid1=001&oid=001&aid=0007980956
10. Available at: www.seoul.co.kr/news/newsView.php?id=20160628030004
11. Available at: www.seoul.co.kr/news/newsView.php?id=20160628030004
12. Available at: http://dh.aks.ac.kr/jsg/index.php/(2016SHWJA)_%EC%97%AC%EB%A6%84_%ED%95%9C%EB%AC%B8_%EC%9B%8C%ED%81%AC%EC%83%B5_Summer_Hanmun_Workshop

    http://dh.aks.ac.kr/Encyves/wiki/index.php/Presentation
13. That digital humanities requires collaboration has become a cliché. The laboratory ideal is enshrined, for example, in "Digital Humanities Manifesto 2.0": available at: http://humanitiesblast.com/manifesto/Manifesto_V2.pdf
14. Available at: kiss.kstudy.com/dataReport/data1.asp
15. Available at: www.dbpia.co.kr/
16. Available at: www.riss.kr/analytics/currentState.do
17. Hung (2003) covers this topic in detail.

References

Bloch, M. (1953/1941), *The Historian's Craft*, Knopf, New York, NY.

Bryant, D. (2008), *The Great Recreation: Ho Ching-Ming (1483-1521) and His World*, Brill, Leiden.

Cha, J. (2015), "Digital/humanities: new media and old ways in South Korea", *Asiascape: Digital Asia*, Vol. 2 Nos 1/2, pp. 127-148.

Chang, K.S. (2010), "Literature of the early Ming to mid-Ming", in Chang, K.S. and Owen, S. (Eds), *Cambridge History of Chinese Literature: From 1375*, Cambridge University Press, Cambridge, pp. 1-62.

Hung, H.-F. (2003), "Orientalist knowledge and social theories: China and European conceptions of East-West differences from 1600 to 1900", *Sociological Theory*, Vol. 21 No. 3, pp. 254-280.

Jockers, M.L. (2013), *Macroanalysis: Digital Methods and Literary History*, University of IL Press, Urbana, IL.

Kim, B. (2016), "Han'guk tijit'ŏl inmunhak kyoyuk ŭi hyŏnhwang kwa kwaje [the current state of digital humanities education in Korea]", in Kim, H. *et al.* (Eds), *Tijit'ŏl Inmunhak Immun*, Hankuk University of Foreign Studies Press, Seoul, pp. 378-398.

Kim, H. (2012), *Inmun Chŏngbohak ŭi Mosaek [in Search of Humanities Informatics]*, Puk K'oria, Sŏngnam.

Kim, H. *et al.* (2016), *Tijit'ŏl Inmunhak Immun [Digital Humanities: A Primer]*, Hankuk University of Foreign Studies Press, Seoul.

Ladurie, E.L. (1979/1968), "The historian and the computer", *The Territory of the Historian*, University of Chicago Press, Chicago, IL, pp. 3-6.

Ong, C.W. (2016), *Li Mengyang: North-South Divide and Literati Learning in Ming China*, Harvard University Asia Center, Cambridge, MA.

Rho, K.H. (2015), *17-segi Chŏnban'gi Han Chung Munhwa Kyoryu [Korean-Chinese Cultural Interactions in the Early Seventeenth Century]*, T'aehaksa, P'aju.

Wagner, E.W. (1963/1971), *Elementary Written Korean*, Harvard-Yenching Institute, Cambridge, MA.

Corresponding author
Javier Cha can be contacted at: javiercha@snu.ac.kr