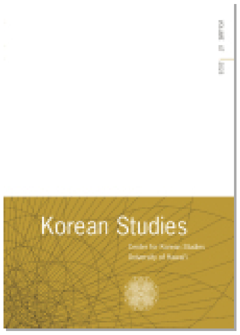
Big Data Studies: The Humanities in Uncharted Waters

Javier Cha

Korean Studies, Volume 47, 2023, pp. 274-299 (Article)

Published by University of Hawai'i Press
DOI: *https://doi.org/10.1353/ks.2023.a908625*



Big Data Studies: The Humanities in Uncharted Waters

Javier Cha

This article discusses the formidable challenges that the advent of big data brings to the digital humanities broadly and proposes some ways the Korean studies community can prepare to navigate these uncharted waters. Standard digital humanities training in data mining, text analysis, mapping, network science, and machine learning will be developed and refined over the coming years, as will research concerning the ephemeral nature of new media, web archives, and the ethics of artificial intelligence. Yet I contend that established responses to the digital transformation of the humanities, while timely and necessary, will prove inadequate for handling petabyte- and exabyte-scale born-digital sources. In the Zettabyte Era, more data is processed in real time than all of the records produced from early times to the 2010s. To make sense of the current information regime, we need critical reflections and comparisons to the classical internet age of the 1990s, the personal computer revolution of the 1980s, and early modern print cultures. This exercise will allow us to situate the humanities in an age of big data as an extension of traditional humanities research and at the same as something foreign.

Keywords: big data studies, big data, humanities, 3Vs, digital materiality

Javier Cha is Assistant Professor of Digital Humanities in the Department of History at the University of Hong Kong, Hong Kong SAR (javiercha@hku.hk).

The Humanities in the Zettabyte Era

In 2012, the Zettabyte Era began. To store one zettabyte on Blu-ray discs, each with a storage capacity of 25 gigabytes and a thickness of 1.2 millimeters, 37.2 billion discs must be stacked 45,000 kilometers high. In 2020, global data production surpassed 59 zettabytes.[1] The hypothetical Blu-ray tower for 59 zettabytes would span 2,637,000 kilometers, or 6.9 times the distance to the Moon.

The astronomical amounts of data being generated in the twenty-first century spur humanists to reflect on the directions, principles, assumptions, and methodologies of our respective disciplines. Standard digital humanities training in data mining, text analysis, mapping, network science, and machine learning will continue to be relevant, as will research concerning the ephemeral nature of new media, web archives, and the ethics of artificial intelligence. Conversely, the new humanities in the Zettabyte Era will require a comprehensive rethinking of everything from material bibliography to data authentication and analytics, archival preservation, and environmental impact. In 2000, John Unsworth introduced the notion of "scholarly primitives" in humanities computing.[2] Twenty-three years later, research practices and pedagogy need to account for the implications of big data's architectural characteristics, as encapsulated in Doug Laney's famous 3Vs: volume, velocity, and variety.[3]

Each of the 3Vs of big data presents unique challenges for the humanities that cannot be dismissed as mere historical repetition. Koreanists, for instance, must debate and determine the boundaries of the Korean web and the portions that should be earmarked for long-term preservation. Mapping the geographic, linguistic, and legal boundaries of the internet, which has never been more dynamic, is a complex task. Every day, millions of South Koreans access digital contents provided by Netflix and YouTube, multinational tech giants headquartered in the United States, through content-delivery networks located on or near the Korean peninsula. Namu Wiki, one of the most popular Korean-language knowledge bases, is operated by Paraguay-based limited liability company Umanle using servers located in Slovakia to partially circumvent South Korea's online censorship laws. Moreover, we must pay attention to new data types. By size, eighty percent of big data is said to be unstructured, and the proportion of the more traditional tabular and textual data produced has decreased over time.[4] How do we archive and navigate the ocean of data that consists primarily of images, audio, video, and 3d point clouds? What about massively multiplayer online role-playing games or blockchainbased metaverses? The third V of big data introduces its own set of difficulties.

This article discusses the formidable challenges that the advent of big data brings to the digital humanities broadly and proposes some ways the Korean studies community can prepare to navigate these uncharted waters. The commendable range of digital humanities approaches demonstrated in this special section will be developed and refined over the coming years. Yet I contend that established responses to the digital transformation of the humanities, while timely and necessary, will prove inadequate for handling petabyte- and exabyte-scale born-digital sources. In the Zettabyte Era, more data is processed in real time than all of the records produced from early times to the 2010s. To make sense of the current information regime, we need critical reflections and comparisons to the classical internet age of the 1990s, the personal computer revolution of the 1980s, and early modern print cultures. This exercise will allow us to situate the humanities in an age of big data as an extension of traditional humanities research and at the same as something foreign.

The Materiality of Big Data

Insofar as digital humanists are concerned with static data sets that can be handled in a single personal computer, big data may not appear to be entirely unprecedented. Upon taking note of the facilities and infrastructure that connect millions of servers around the world, however, we realize that big data's characteristics are distinct from those of previous information regimes and media landscapes. In 2014, Facebook processed 4 PB[5] and Naver 1.2 billion requests for its 18 million portal users daily.[6] With 50 billion photos posted to its servers as of 2022, Instagram is the largest repository of visual materials ever created.[7] The total amount of video uploaded to YouTube from 2005 or 2019 is 95,000 years.[8] In addition to search portals and social media platforms, 6.6 billion smartphones,[9] 1.5 billion personal computers,[10] tens of millions of servers,[11] and countless sensors contribute to the global production of big data. As one among 2.9 billion monthly active users of Facebook and more than 1 billion users of Google Drive, I uploaded 4.7 GB and 1.05 TB to their servers, respectively.

The *Oxford English Dictionary* defines big data succinctly as "data of a very large size, typically to the extent that its manipulation present significant logistical challenges."[12] The casual use of the buzzword "big data" to refer to any ostensibly large data set, or as a substitute word for data science, disregards the significance of data materiality.[13] While extremely large databases introduce new analytical problems, query services that operate at a high abstraction layer, such as Google BigQuery and Amazon Athena, eliminate most of these difficulties. Most pertinent to the future direction of the humanities is what the *Oxford English Dictionary* means by "significant logistical challenges." Humanists are trained not to treat documents, books, inscriptions, and databases as self-evident stores of information. Early modern historians pay close attention to printing technologies, paper production, ink, circulation, reception, and reading cultures when handling primary sources. Similarly, source criticism in the context of big data entails engagement with the physical infrastructure designed to enable the handling and circulation of petabytes and exabytes of data: data centers.

Given that big data exists as a distributed collection of digital storage devices, it is essential to describe the hardware required to store and process massive amounts of data. In 2023, a typical notebook computer has one terabyte of storage in the form of an M.2 NVMe solid-state drive that is 22 mm wide and 80 mm long. One petabyte is equivalent to 1024 one-terabyte M.2 drives, while one exabyte constitutes 1024 PB. A data migration service offered by Amazon Web Services (AWS) helps us visualize the physicality of one exabyte, or 1,048,576 terabytes. The transmission of 1 EB over a 10 Gbps fiber optic connection is estimated to take twenty-six years.[14] AWS Snowmobile "loads" extremely large amounts of data onto semi-trailer trucks carrying shock-proof hard disks; each truck is capable of transporting 100 PB, allowing the transfer of exabyte-scale data in weeks rather than decades.[15]

At the exabyte scale, our analytical bibliography of big data turns to supersized data centers at multiple locations around the world built by cloud industry leaders such as Facebook, Amazon, Microsoft, Google, Alibaba, and Naver. Each facility is the size of several football stadiums equipped with hundreds of thousands to millions of hard disks and solid-state drives. Facebook operates a 150,000-m^2 data center cluster in Clonee, Ireland, which as of 2019 consisted of three facilities and ten more under construction. In 2014, Naver launched its flagship data center Kak (각 or 閣 as a tribute to Changgyŏnggak 藏經閣, which housed the *Tripiṭaka Koreana*) in Ch'unch'ŏn, with 120,000 servers capable of handling more data on its first day than "ten-thousand National Libraries of Korea combined."[16]

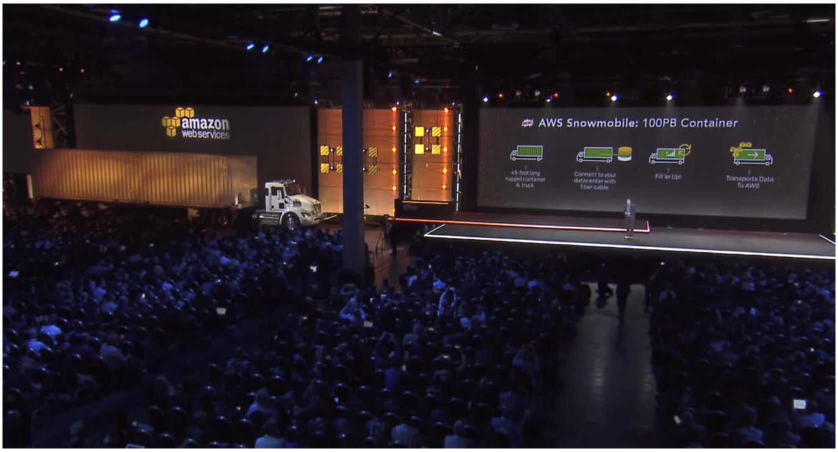

amazon web services
AWS Snowmobile: 100PB Container
45-foot long rugged container & truck
Connect to your datacenter with fiber cable
Fill 'er Up!
Transports Data To AWS

Fig. 1. AWS Snowmobile. Source: https://ind01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fyoutu.be%2F8vQmTZTq7nw%3Ft%3D123&data=05%7C01%7Cpraveen.s%40thomsondigital.com%7C506d4976760d492304cd08db8e33ee07%7Cbfe0633a6c7947d5b23e9aea466111c7%7C0%7C0%7C638260125922619388%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=UuJM63C%2Bj0nFKqnGOeKbvReY7lHA3Sp2madEh0TWwwQ%3D&reserved=0

Due to the massive physical infrastructure required to store and process modern-day primary sources, energy has emerged as another "significant logistical challenge" of big data. In the context of a small number of personal computers and mobile devices, the fact that digital media, unlike paper, requires electricity to store and retrieve information is not particularly problematic. Hundreds of server units have already begun to cause issues, however. In 2008, the National Library of Korea's opening of the new digital library placed a tremendous strain on its electrical grid, necessitating the installation of a new substation and a dedicated power source.[17] Data centers have substantially higher energy needs. In 2020, Naver's Kak used 156,875 MWh of electricity.[18] Facebook's Clonee site has access to 642 MW, which is enough to power 300,000 average American homes.[19] As of 2020, data centers consume approximately one percent of the world's electricity, and this figure is expected to increase in the coming decades.[20]

Meanwhile, global data production continues to grow at an unprecedented rate. A 2012 prediction anticipated that 40 zettabytes of data would be generated by 2020,[21] but the actual amount turned out to be 59 zettabytes. According to a 2018 report, 175 zettabytes will be created by 2025.[22] Considering that the Covid-19 pandemic caused a 47 percent increase in internet usage in 2020, the total size of global big data in the 2020s is anticipated to surpass all expectations.[23]

Archiving Big Data

The 1493 Korean reprint of the Chinese encyclopedia *Gujin shiwen leiju* 古今事文類聚 (Korean *Kogŭm samun yuch'wi*) was one of the most prized books of its time. In 2008, as a graduate student attending a seminar on book history, I discovered that one copy was available in the university's rare book collection. Despite being more than 500 years old, the volume that my classmates and I examined was in pristine condition, which was perhaps not surprising given the circumstances of its publication. King Sŏngjong 成宗 (r. 1469–1494), who personally commissioned the project that took eight years to complete, ensured that the books were printed on high-grade paper.

The same cannot be said about digital sources. Most consumer-grade digital storage media are not capable of retaining data for 500 years. Data in

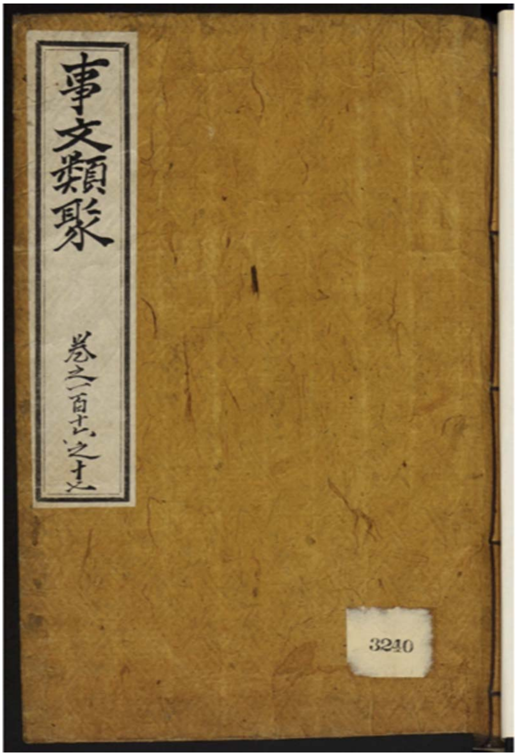


Fig. 2. The 1493 Korean reprint of the Chinese encyclopedia *Gujin shiwen leiju*. Courtesy of Harvard-Yenching Library's Korean rare book digitization project https://iiif.lib.harvard.edu/manifests/view/drs:9345235$5i.

CPU cache and RAM are lost immediately without power. The lifespans of disks depend on whether they experienced prolonged stress, such as constant reads and writes in a server, and whether they were designed for performance or archival purposes. Whereas solid-state drives are relatively more resistant to physical shocks, they are said to retain data typically for about 7 to 10 years on average; while hard disks generally last for up to 30 years, helium-sealed enterprise-grade drives are less likely to experience failure and corruption. Archival-grade optical discs are estimated by Kodak to be able to guarantee data integrity for up to 200 years,[24] but only under specific conditions. The discs must be kept vertically to prevent them from sticking to the case and held in a climate-controlled room set to 25 degrees Celsius and 40% humidity.[25]

The foregoing discussions of data volume, data centers, and energy consumption are significant to the extent that big data can be preserved for posterity. How will future historians study the 2010s and 2020s? What will archaeologists be able to unearth from the ruins of Kak? The ephemeral nature of digital media has reversed our relationship with primary sources. On this ground, archivists and technologists have voiced concerns about the digital dark age since the 1980s, which has largely gone unaddressed and has been exacerbated by the world wide web and big data. The digital historian Roy Rosenzweig initiated the abundance versus scarcity debate in response to the emergence of inexpensive mass storage solutions such as optical discs and Web 1.0 hypermedia.[26] Web 2.0 has further complicated this. Even setting aside the problem of bit durability, the physical presence of big data in the form of data centers and global communications infrastructure means that decades from now, let alone centuries, people will not have the option of excavating what remains of today's digital devices and servers, powering them up, and extracting information from them.

Our access to the past web relies almost exclusively on the Internet Archive. Between 1996 and 2020, the San Francisco-based nonprofit saved 733 billion web contents totaling 70 PB to its digital archive the Wayback Machine, and it continues to collect more items.[27] Among the many collections saved in the Wayback Machine is the Web 1.0 community GeoCities, which had 38 million homepages from 1994 to 2009.[28] Ian Milligan's research on this early internet community[29] was made possible thanks to the Wayback Machine. While the Wayback Machine is an invaluable resource, 70 PB represent a marginal portion of the 59 zettabytes of total global data production and less than the storage capacity of a single AWS Snowmobile truck. Most of the preserved content, moreover, is currently accessible as an uncatalogued data dump, which has yet to be fully indexed and checked for geographic and linguistic distribution.[30] Among the Internet Archive's "catalogued" contents, an advanced search for Korean-language materials returned 66,308 results on 9 May 2022; the equivalent search for English returned 36,683,156 hits.

The archiving of massive amounts of data places a tremendous burden on non-profit organizations and public institutions. In 2010, the U.S. Library of Congress attempted to archive Twitter. This ambitious project was conducted in collaboration with the social media aggregator Gnip and Twitter headquarters, which provided public tweets from 2006 to 2010.[31] Until December 2012, the Library of Congress archived the content and metadata of 150 billion tweets worth 132 terabytes and continued to receive more data.[32] The Twitter archive project received considerable media attention and inquiries from more than 400 researchers around the world.[33] Unfortunately, the library was unable to develop and launch a viable search engine for it. A report published in January 2013 noted that "executing a single search of just the fixed 2006–2010 archive on the Library's systems could take 24 hours."[34] This anecdote serves as a reminder of the amount of human, financial, and technological resources required to return instant results on Twitter, which the archival version of it was unable to provide. In 2017, the Library of Congresses announced the project's premature termination.[35]

The preservation of big data over the long term requires public–private partnerships with tech giants, but the viability and specifics of such endeavors are largely unknown. What will occur if Amazon, Microsoft, or Facebook cease to operate their data centers? Will the data stored in infrastructure managed by Google, Alibaba, or Naver remain secure and accessible in the future? Naver's promotional slogan created during the launch of Kak reads, "We protect what you leave behind. We pass on the records of today to tomorrow."[36] Is this really the case? In contrast to the goal that "the data created by Naver users must be handed down to future generations in perpetuity,"[37] the archiving of big data is a complex problem constrained by technology and privacy laws. Consider the following from Naver Kak's official data preservation policy:

> Just as not all services are duplicated, not all data is backed up, nor are all backup copies retained indefinitely. In practice, the majority of servers in the data center are used for service, while only a smaller portion are used for backup. As required by applicable laws and ordinances, transaction, payment, and personal information logs are only kept for five years, but the remainder of the services are retained based on the underlying philosophy of the service. The goal for personal data stored in services that require a Naver login, such as personal

> emails, blogs, cafes, and Ndrive, is to keep them forever. This is consistent with our company's guiding philosophy that individual records constitute the records of the present. Except for specialized services such as [Naver] News Library, the logs of regular newspaper articles, advertisements, and Knowledge iN searches that do not require a login are deleted after one week. In situations where backup is ineffective due to long replication times resulting from a large data capacity or number of files, we rely on data redundancy across data centers as a substitute.[38]

Naver's meteoric rise as one of the largest corporations in South Korea owes much to the transition to personalized services during the Web 2.0 era, in which companies generate profit from subscription fees and advertising revenue. In contrast to the digital edition of the *Annals of the Chosŏn Dynasty* (*Chosŏn wangjo sillok* 朝鮮王朝實錄) or the GeoCities web archive, Naver's online services, and the infrastructure that enables them, are not designed with the public interest in mind. To be fair, Naver's mission statement acknowledges the extent to which the company's vast user data and internal logs represent contemporary South Korea's living records. When data no longer contributes to the company's bottom line, a private enterprise does not have the responsibility to "pass on the records of today to tomorrow."

On numerous occasions, online services have deleted or lost data that users believed would last for a long time. Flickr, a photo-sharing platform, deleted the images of its free users en masse in 2019 due to its parent company's financial troubles.[39] Due to a faulty server migration process, Myspace accidentally "lost every single piece of content uploaded to its site

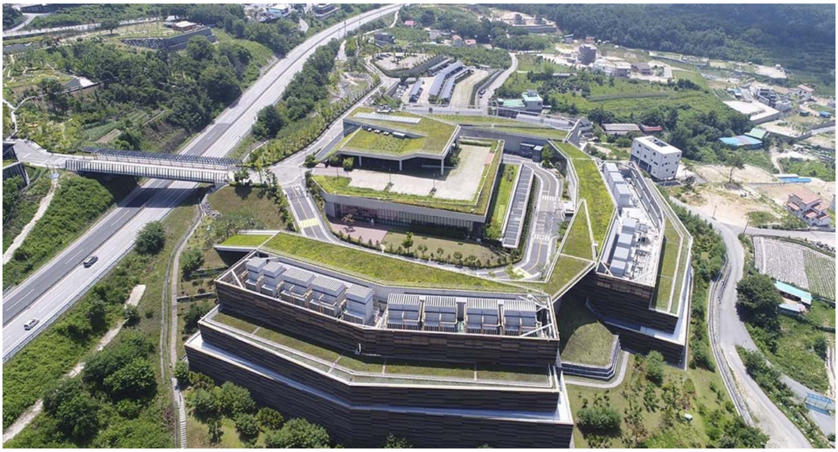


Fig. 3. Naver's data center Kak.

before 2016."[40] The retention of non-profitable data represents a significant burden even—or especially—for large corporations. In 2020, twenty-one years after its founding, the South Korean social media platform Cyworld, which at its peak had 32 million users and more than 100 billion won in annual revenue from online product sales, was on the verge of shutting down. Upon learning that Cyworld's servers would soon go offline, users turned to the programmer O Kilho's CyBackup, which allows users to create personal archives of their activities and photos shared on the platform.[41] Fortunately, an eleventh-hour capital injection allowed the resumption of services, but any user content that had not been saved ran the risk of being lost forever. As evidenced by these preceding examples, data preservation is costly. In 2019, Russia's leading telecommunications firms sought government subsidies for the additional equipment purchases required to comply with the data surveillance provisions of the Yarovaya amendments.[42] It was estimated that the legal obligation to store call and message content for six months and metadata for three years would set each operator back $627 million US dollars, or 40 billion Russian rubles, over a five-year period.

Beyond just volume, the distributed and dynamic nature of big data complicates the traditional definitions of an archive. As Naver's Kak demonstrates, big data can be replicated, marked for deletion, retained briefly, or preserved for an extended period. Approximately 90% of global data production consists of redundant copies.[43] When a South Korean user watches Squid Game on Netflix, the video is not transmitted directly across the Pacific Ocean from Los Gatos, California, via submarine fiber optic cables. Instead, content for the South Korean market is stored and maintained locally. This complex procedure involves cached libraries hosted in AWS S3 (Simple Storage Service) and a content-delivery solution known as Open Connect, which installs Netflix-specific servers directly within internet service provider facilities.

According to traffic demands, Web 2.0 platforms categorize data as hot, warm, and cold, and manage them differently. At Facebook, for example, the most recent news and profile information that must be processed immediately upon login are placed in hot storage, whereas old information that users rarely access is kept in cold storage. From the early years of service, Facebook's challenge has been the management of binary large objects (BLOBs) such as photos and videos. According to information disclosed in 2013, 82% of Facebook traffic was used to transmit just 8% of photo data.[44] To store and handle the remaining 92% of BLOB, Facebook built a special cold storage facility system called

Haystack on a 370,000-m^2 site in Prineville, Oregon, a small city with a population of only 9000.[45] In 2013, Facebook's cold storage consisted of two-petabyte racks consisting of 500 hard disks in RAID-6 arrangement for added stability and capable of reducing the power consumption of units with low data demand.[46]

Engineers have devised innovative solutions, focusing primarily on bit rot, to address the challenges of storing digital materials for the very long term. In 2009, Millenniata developed the Millennial Disc (M-Disc), an optical disc capable of retaining data for one thousand years. M-Disc is specially designed to resist rust and maintain data integrity under extreme conditions.[47] In 2011, the United States Department of Defense conducted a stress test that expose archival-grade optical discs to a broad spectrum of light for 24 hours in an environment held at 85 degrees Celsius and 85 percent humidity.[48] Only M-Disc did not experience any data loss.[49] Microsoft Research's Project Silica, another promising archival medium, uses femtosecond laser to inscribe data on a 2-mm fused silica glass.[50] Project Silica is officially rated to last ten thousand years; according to an interview my lab conducted in August 2020 with Ant Rowstron, the Deputy Lab Director, the actual lifespan could exceed one million years.

The Third V: Variety

The move from old to new media, from Web 1.0 to 2.0, and from personal computing to the cloud is a complex and non-linear transformation. To a certain extent, the situation we face today bears similarities to what historians have demonstrated about the transition from manuscript to print culture in Europe, East Asia, and other regions. In this vein, Robert Darnton compared the early modern "anecdotes" written by "paragraph men" to the fragmentary nature of blogging and the social web.[51] Matthew Kirschenbaum's innovative application of digital forensics to English literature and media studies has striking parallels to studies of paratext, marginalia, and hidden layers in premodern artifacts.[52]

However, I would caution against overstating the parallels and continuities with superficially similar past phenomena, and I reckon that Darnton and Kirschenbaum would maintain the same. The resource-intensive infrastructure that houses contemporary primary sources is unlike anything historians and archivists have encountered. By definition, big data does not reside in a single location; unlike static data sets used in conventional digital humanities research, big data cannot be traced to a specific storage device. Extremely large databases are partitioned into thousands of shards and stored on multiple servers, and data streams self-replicate and migrate in response to user activities, dynamically forming new networks and boundaries.

Variety adds another layer of complexity to the modern information regime driven by big data. Typically, the term "data" conjures up an image of a matrix containing text and numbers, which can be thick or thin depending on the number of columns and rows. When the tabular arrangement proves to be cumbersome, a relational database is created by separating and organizing redundant entries into multiple tables. Due to the social and participatory nature of Web 2.0, engineers have proposed and developed new kinds of database management systems. Graph databases, for example, provide native support for managing entities and relationships and store records as node and edge properties. However, the data explosion of the past decade was not necessarily comprised of structured text and numeric data or nodes and edges. As previously mentioned in the discussion of Facebook's cold storage mechanism, the proportion of unstructured BLOBs has surged during the Web 2.0 era and new types of binary objects are being introduced to digital ecosystems.

Smartphones are primarily responsible for this sea change. As of 2022, there are 6.6 billion smartphone users worldwide,[53] including 95% of South Korean adults and 76% of those who live in advanced economies.[54] Smartphones are intelligent devices equipped with an array of advanced sensors, including cameras, lenses, gyroscopes, compasses, location trackers, and depth sensors. Consider what happens when a smartphone user takes a photograph. Depending on the settings, a semiprofessional-grade CMOS sensor captures the subject's light via the main, ultrawide-angle, or telephoto lens. Internally on the device, substantial post-processing occurs. Software enhancers may use ToF or LiDAR sensor-collected three-dimensional data to render the photograph into a more visually appealing form. Advanced triangulation techniques that combine data from GPS satellites, cell towers, and WiFi routers yield a precise approximation of the geocoordinates where the user captured the image, which is embedded in the file.

Visual materials are not novel in the humanities, and, privacy concerns aside, automated location tagging is a welcome feature. However, social media, wearable devices, and virtual reality worlds also collect vast quantities of BLOBs that are unfamiliar to humanities researchers. To facilitate what Shoshana Zuboff has termed "surveillance capitalism,"[55] for example, Web 2.0 services enable highly intrusive trackers to collect information about user behavior and psychology. Facebook offers its advertising partners the Pixel service, in addition to processing user-uploaded contents for targeted advertising. When a few lines of JavaScript code are inserted into a third-party website, information regarding the users' online activities outside of Facebook is gathered and sent to Facebook's server. The profiling is based on not solely the content, but also on patterns of interaction with the interface, such as clicks, taps, cursor movements, and pauses. And Facebook is not alone in tracking its users' behavior. On content management systems, such as WordPress, a number of plugins offer similar functionalities. The South Korean startup Four Grit provides a user experience analytics service called Beusable that collects and visualizes online customer behavior data.[56]

Biometric data are another uncharted territory. Both Android and iOS device users increasingly unlock their devices with facial or fingerprint recognition. The latest versions of wearable devices from Fitbit, Apple, Garmin, Huawei, Samsung, and others monitor the user's sleep patterns, physical activities, pulse, and blood oxygen saturation. The performance of these biometric sensors is adequate for some medical professionals to use them as a reference. For instance, a research team at the University of California, San Francisco, proposed a method for combating COVID-19 using health data from wearable devices.[57] Should private information such as browsing habits, psychological profiles, and biometric records be made accessible to researchers? If so, what insights can humanists glean from such data? I am not strongly arguing for or against preserving any sensitive data for future generations. However, I do believe it is necessary to have constructive discussions about creating secure, responsible, and sustainable archives of various types of data that may one day prove useful.

South Koreans have become accustomed to encountering buzzwords such as the Fourth Industrial Revolution, interdisciplinary convergence, artificial intelligence, the metaverse, smart cities, and the internet of things. While the specific keywords and slogans change according to presidency and fad, they share a common thread. Due to technological advances, it has become possible to collect visual, auditory, tactile, olfactory, and gustatory data, even though the detection and datafication of some sensory information are more refined than others. Beyond the five senses, intelligent electronic devices also collect a vast amount of data that is not perceptible to humans. In the years ahead, the line between virtual and physical worlds will continue to blur. In 2020, Naver created an incredibly detailed three-dimensional map of Seoul in which each pixel represented eight centimeters.[58] Using advanced photogrammetry, the model stitched together 25,463 aerial photographs covering 605.2 km^2 and 600,000 buildings over a thirty-day period. The goal was to assist Naver in preparing for autonomous driving, which will collect an even greater volume of data about Seoul once it is commercially available. In 2019, approximately 5 million cars traversed the streets of Seoul.[59] If these vehicles are replaced with autonomous ones, Seoul could gain millions of data collectors that sense 140 MB of data per second about their surroundings.[60] Should the orthomosaic maps or millisecond-level snapshots of Seoul be regarded as virtual representations of the city? Or as Seoul itself? To address these questions indirectly, I invite readers to consider a surprising example: IKEA catalogs. The Swedish furniture company, whose products are adored by a large number of South Koreans, entices prospective customers with displays of appealing Nordic minimalist design. What few people realize is that more than 75% of product images in the IKEA catalog are computer-generated.[61]

Conclusion

In 2010s and 2020s, the surge of interest in computational and digital methods in Korean studies has been astounding.[62] In December 2018, *Korean Historical Review* (*Yŏksa hakpo* 歷史學報) published six articles surveying the current state of digital historical scholarship worldwide. The annual meeting of the Association for Asian Studies in 2022 included a roundtable discussion on digital humanities in relation to Korean studies. In April 2022, KAIST officially launched the School of Digital Humanities and Computational Social Sciences and announced the ambitious goal of hiring digital specialists to fill half of new faculty positions over the next five years.[63] This special section of *Korean Studies*, devoted to digital Korean studies, contributes to this ongoing trend.

The question is what digital humanities means in the Korean context. According to a 2017 survey of South Korean scholars in social science and humanities, ninety-four percent believe that "digital humanities methodology" is necessary for their field.[64] Only four percent of respondents, however, indicated that they used computational methods (referred to in the survey as "statistical methods") such as topic modeling and text mining in their research, even though twenty-three percent have received training in "big data analysis method."[65] Sixty-four percent associate digital humanities with the web-based access to primary and secondary sources. Sixteen percent use Google Maps or Google Earth to visualize spatial data, and another sixteen percent use digital technology as a platform for social media engagement. Will training the next generation of humanists with Google Earth, Python, network analysis, and other data-driven or data-assisted methods be sufficient?

Robert Darnton's conclusion in his opinion piece on blogs provides some hints: "I don't believe that history teaches lessons, at least not in a direct, easily applied manner, but it does raise questions."[66] The big data turn affords area studies specialists an opportunity to evaluate the continued relevance of what has for long a time been an open and interdisciplinary field. The digital approach to the fuzzy notion of Korea includes explorations of the core digital South Korea, North Korea's distinct information technology sectors discussed in Benoit Berthelier's contribution, and the Korean diaspora. As the concept of data sovereignty gains prominence, experts on Korea's national identity and nationhood will want to investigate its fascinating digital layer, which includes data centers, telecommunication networks, electric power grids, and online participants of the Korean web—the definitions of which will be highly contested. Additionally, comparisons will generate new research topics. The European Union enables the uninterrupted flow and sharing of user data and electricity among its member states, whereas East Asia lacks such an arrangement. What are the implications and consequences of this difference? To create archives of the vast amounts of data currently being generated in South Korea, we will need to foster public engagement and citizen participation while respecting and adhering to the country's data protection and privacy laws.

Finally, I would like to propose that the digital humanities be kept distinct from its sister disciplines such as computational social sciences and cultural data science. An abundance of topics at the crossroads of big data and the humanities calls for the insights of Korea experts. I hope that my fellow Koreanists refrain from reflexively equating the big data turn in the humanities with quantitative methodologies or data science, although some degree of familiarity with analytics is crucial for nurturing digital literacy. My recommendation is that we, as humanists, study big data in a manner that builds on our strengths in linguistic proficiency, historical understanding, ethnographic inquiry, and critical thinking, rather than following the paths of scientists and engineers. This should be done while cultivating a harmonious relationship with our technically oriented colleagues and potential collaborators.

Notes

1. International Data Corporation, “IDC’s Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data,” May 8, 2020, https://www.idc.com/getdoc.jsp?containerId=prUS46286020.

2. John Unsworth, “Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?” (Paper presented at the Symposium on Humanities Computing: Formal Methods, Experimental Practices, King’s College, London, May 13, 2000), https://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

3. The Gartner Group is generally credited with introducing the notion of 3Vs. Doug Laney coined the terms data volume, data velocity, and data variety, collectively known as the 3Ds, in a report submitted to his employer, Meta Group, years before the rise of the cloud industry. In 2004, Meta Group was acquired by Gartner, hence the confusion surrounding its proper attribution. See Doug Laney, “3D Data Management: Controlling Data Volume, Velocity, and Variety,” *META Delta*, February 6, 2001, https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

4. Michael O’Reilly, “The Unseen Data Conundrum,” *Forbes*, February 3, 2022, https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/.

5. Janet Wiener and Nathan Bronson, “Facebook’s Top Open Data Problems,” *Facebook Research*, October 22, 2014, https://research.facebook.com/blog/2014/10/facebook-s-top-open-data-problems/.

6. Naver Business Platform, *Teit’ŏ sent’ŏ Kak* (Seoul: Iro, 2015), 18.

7. Darren Bernhardt, “Instagram Obsessed: Can We Vacation Without Posting Every Moment?,” *CBC News*, June 29, 2019, https://www.cbc.ca/news/canada/manitoba/instagram-vacation-sharing-social-media-1.5187677.

8. J. Clement, “Hours of Video Uploaded to YouTube Every Minute as of May 2019,” *Statista*, August 20, 2020, https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

9. Statista Research Department, “Number of Smartphone Subscriptions Worldwide From 2016 to 2021, With Forecasts From 2022 to 2027,” *Statista*, August 22, 2022, https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

10. Statista Research Department, “Installed Base of Personal Computers (PCs) Worldwide From 2013 to 2019,” *Statista*, September 15, 2016, https://www.statista.com/statistics/610271/worldwide-personal-computers-installed-base/.

11. Thomas Alsop, “Server Shipments Worldwide From 2010 to 2020,” *Statista*, November 28, 2022, https://www.statista.com/statistics/219596/worldwide-server-shipments-by-vendor/.

[12] OED Online, "Big Data," accessed August 1, 2020, https://www.oed.com/view/Entry/18833#eid301162178.

[13] For example, Frédéric Kaplan, "A Map for Big Data Research in the Humanities," *Frontiers in Digital Humanities* 2, no. 1 (2015): 2, https://doi.org/10.3389/fdigh.2015.00001 references the same Oxford English Dictionary entry mentioned in this article but disregards the logistical aspects of big data. Similarly, digital humanities and information studies scholars tend to employ a vague definition of big data as massive data sets without specifying the volume; velocity and variety are overlooked. See Amalia S. Levi, "Humanities 'Big Data': Myths, Challenges, and Lessons," in *2013 IEEE Conference on Big Data* (Silicon Valley, CA: IEEE, 2013), 33–6, https://doi.org/10.1109/BigData.2013.6691667 and Richard Hawkins, "Use of Big Data in Historical Research," in *Big Data in the Arts and Humanities: Theory and Practice*, ed. Giovanni Schiuma and Daniela Carlucci (Boca Raton, FL: CRC Press, 2018), 77–87.

[14] Tony Yoo, "Amazon's New Tool for Startups is a Massive Truck," *Business Insider Australia*, December 1, 2016, https://www.businessinsider.com.au/amazons-new-tool-for-data-hungry-startups-is-a-massive-truck-that-drives-to-the-cloud-2016-12.

[15] Amazon Web Services, "AWS re:Invent 2016: Move Exabyte-Scale Data Sets with AWS Snowmobile," December 1, 2016, https://youtu.be/8vQmTZTq7nw.

[16] Naver Business Platform, *Teit'ŏ sent'ŏ Kak*, 34.

[17] Pak Chinho, personal communication, November 14, 2022 and email message to author, December 7, 2022.

[18] Naver, "DATA CENTER GAK," accessed November 15, 2022, https://datacenter.navercorp.com/green/green-energy.

[19] Rory Carroll, "Why Ireland's Data Centre Boom is Complicating Climate Efforts," January 6, 2020, https://www.irishtimes.com/business/technology/why-ireland-s-data-centre-boom-is-complicating-climate-efforts-1.4131768.

[20] On how to calculate the global total power consumption of data centers, experts are divided. Some estimates only account for the energy required to operate data center facilities, while others include telecommunication networks. In 2014, 1.8% of all electricity generated in the United States was consumed by data centers. See Arman Shehabi et al., *United States Data Center Energy Usage Report* (Berkeley, CA: Lawrence Berkeley National Laboratory, 2016), ES-1, https://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf. According to Eric Masanet et al., "Recalibrating Global Data Center Energy-Use Estimates," *Science* 367, no. 6481 (February 2020): 984–6, https://doi.org/10.1126/science.aba3758, data centers consume 1% of all electricity generated worldwide.

In 2017, the British newspaper *Guardian* published an editorial by Climate Home News that claimed a "tsunami of data" would devour "one-fifth of global electricity by 2025." See Climate Home News, "'Tsunami of Data' Could Consume One Fifth of Electricity by 2025," *Guardian*, December 11, 2017, https://www.theguardian.com/environment/2017/dec/11/tsunami-of-data-could-consume-fifth-global-electricity-by-2025. This estimate was derived by summing the power requirements of all hardware that connects to the internet or enables online services, such as smartphones, smart TVs, security cameras, servers, fiber-optic cables, and cellular towers. It was also based on the assumption that no significant efficiency gains will be observed by 2025, despite contrary industry reports and trends. In 2015, a Huawei research team from Sweden simulated the absolute worst case, in which data centers consume 51% of electricity by 2030. See Anders S. G. Andrae and Tomas Edler, "On Global Electricity Usage of Communication Technology: Trends to 2030," *Challenges* 6, no. 1 (2015): 137, https://doi.org/10.3390/challe6010117. This scenario is extremely unlikely to come true.

21. John Gantz and David Reinsel, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East* (Framingham, MA: International Data Corporation, 2012), 1.

22. David Reinsel, John Gantz, and John Rydning, *Data Age 2025: The Digitization of the World from Edge to Core* (Framingham, MA: International Data Corporation, 2018), https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

23. OpenVault, *OpenVault Broadband Industry Report (OVBI): Q1 2020* (Hoboken, NJ: 2020), 2, https://openvault.com/NEW-SITE-OV3/wp-content/uploads/2021/02/Openvault_Q120_DataUsage_FINAL.pdf.

24. Fred R. Byers, *Care and Handling of CDs and DVDs* (Washington, DC: Council on Library and Information Resources, National Institute of Standards and Technology, 2003), 13, and Kodak, "Permanence and Handling of CDs," October 13, 1999, http://web.archive.org/web/19991013135327/https://www.kodak.com/global/en/professional/products/storage/pcd/techInfo/permanence.shtml.

25. Byers, *Care and Handling of CDs and DVDs*, 13 and Kodak, "Permanence and Handling of CDs."

26. Roy Rosenzweig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108, no. 3 (2003): 735–62, https://doi.org/10.1086/ahr/108.3.735.

27. Katie Barrett, "On Preserving Memory," *Internet Archive Blogs*, December 18, 2020, http://blog.archive.org/2020/12/18/on-preserving-memory/.

28. The US version ended the service in 2009. GeoCities Japan remained active until 2019.

29. Ian Milligan. *History in the Age of Abundance? How the Web is Transforming Historical Research* (Montreal & Kingston: McGill-Queen's University Press, 2019).

30. On May 9, 2022, in a private conversation with Mark Graham, director of the Wayback Machine, I learned that the Internet Archive is interested in analyzing the linguistic distribution of its archive data but that the volume of data currently makes this task difficult to execute.

31. Andrew McGill, "Can Twitter Fit Inside the Library of Congress?" *The Atlantic*, August 4, 2016, https://www.theatlantic.com/technology/archive/2016/08/can-twitter-fit-inside-the-library-of-congress/494339/ and Matt Raymond, "How Tweet It Is! Library Acquires Entire Twitter Archive," *Library of Congress Blog*, April 14, 2010, https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/.

32. Library of Congress, "Update on the Twitter Archive at the Library of Congress," January 2013, https://www.loc.gov/static/managed-content/uploads/sites/6/2017/02/twitter_report_2013jan.pdf.

33. Library of Congress, "Update on the Twitter Archive," January 2013, 3.

34. Ibid.

35. Library of Congress, "Update on the Twitter Archive at the Library of Congress," December 2017, https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

36. Naver TV, "Teit'ŏ sent'ŏ 'Kak' kirok yŏngsang," September 27, 2013, 4:02–4:14, https://tv.naver.com/v/86767/list/8547.

37. Naver Business Platform, *Teit'ŏ sent'ŏ Kak*, 14.

38. Ibid, 204

39. Katie Notopoulos, "Flickr Is Deleting Your Photos Soon. Here's How To Save Them," *BuzzFeed News*, January 8, 2019, https://www.buzzfeednews.com/article/katienotopoulos/how-to-save-flickr-photos-deleted-download.

40. Alex Hern, "Myspace Loses All Content Uploaded Before 2016," *Guardian*, March 18, 2019, https://www.theguardian.com/technology/2019/mar/18/myspace-loses-all-content-uploaded-before-2016.

41. O Kilho, "Ssaiwŏldŭ sajin paegŏp," *Kilho.net*, October 2018, https://kilho.net/archives/various/2190. The source code is available at https://github.com/newkilho/CyBackup.

42. Nadezhda Tsydenova, "Russian Telecoms Firms Want Government Compensation for Data Storage Law Costs," *Reuters*, October 29, 2019, https://www.reuters.com/article/us-russia-internet-operators-idUKKBN1X813P.

43. International Data Corporation, "IDC's Global DataSphere."

44. See Krish Bandaru and Kestutis Patiejunas, "Under the Hood: Facebook's Cold Storage System," *Facebook Engineering*, May 4, 2015, https://engineering.fb.com/core-data/under-the-hood-facebook-s-cold-storage-system/ and Rich Miller, "Inside Facebook's Blu-Ray Cold Storage Data Center," *Data Center Frontier*, July 1, 2015, https://datacenterfrontier.com/inside-facebooks-blu-ray-cold-storage-data-center/.

45. Mike Rogoway, "Facebook Plans Ninth Data Center in Prineville, Says Total Spending There Will Top $2 Billion," *The Oregonian*, June 12, 2020, https://www.oregonlive.com/silicon-forest/2020/06/facebook-plans-ninth-data-center-in-prineville-says-total-spending-there-will-top-2-billion.html and Peter Vajgel, "Needle in a Haystack: Efficient

Storage of Billions of Photos," *Facebook Engineering*, April 30, 2009, https://engineering.fb.com/core-data/needle-in-a-haystack-efficient-storage-of-billions-of-photos/.

46. Rogoway, "Facebook Plans Ninth Data Center" and David Rosenthal, "Facebook's Warm Storage," *DSHR's Blog*, October 23, 2014, https://blog.dshr.org/2014/10/facebooks-warm-storage.html.

47. Sebastian Anthony, "M-Disc is a DVD Made Out of Stone that Lasts 1,000 Years," *ExtremeTech*, August 10, 2011, https://www.extremetech.com/computing/92286-m-disc-is-a-dvd-made-out-of-stone-that-lasts-1000-years.

48. Ivan Svrcek, *Accelerated Life Cycle Comparison of Millenniata Archival DVD* (China Lake, CA: Life Cycle and Environmental Engineering Branch, Naval Air Warfare Center Weapons Division, US Department of Defense, 2009), 2–3.

49. Svrcek, *Accelerated Life Cycle Comparison*.

50. Jim Salter, "Microsoft's Project Silica Offers Robust Thousand-Year Storage," *Ars Technica*, November 7, 2019, https://arstechnica.com/gadgets/2019/11/microsofts-project-silica-offers-robust-thousand-year-storage/.

51. Robert Darnton, "Blogging, Now and Then," *The New York Review of Books*, March 18, 2010, https://www.nybooks.com/daily/2010/03/18/blogging-now-and-then/.

52. Matthew G. Kirschenbaum, *Mechanisms: New Media and the Forensic Imagination* (Cambridge: MIT Press, 2008) and *Bitstreams: The Future of Digital Literary Heritage* (Philadelphia: University of Pennsylvania Press, 2020).

53. Statista Research Department, "Number of Smartphone Subscriptions."

54. Kyle Taylor and Laura Silver, "Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally," *Pew Research Center*, February 5, 2019, 3, https://www.pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf.

55. Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (New York: Public Affairs, 2019).

56. Beusable's website is found at https://www.beusable.net/.

57. Ariel Bleicher and Katherine Conrad, "We Thought It Was Just a Respiratory Virus: We Were Wrong," *UCSF Magazine*, Summer 2020, https://magazine.ucsf.edu/we-thought-it-was-just-respiratory-virus.

58. Ch'oe Hansŭng, "Neibŏ ka mandŭrŏ kanŭn chincha kattŭn kasang tosi: HD maep ŭi haeksim, kojŏngmil 3D chido e taehayŏ," September 22, 2020, https://blog.naver.com/naver_diary/222096345088.

59. Seoul Transport Operation & Information Service, *2019 Sŏul t'ŭkpyŏlsi kyot'ongryang chosa charyo* (Seoul: Seoul Metropolitan Government, 2019), 6, https://news.seoul.go.kr/traffic/files/2020/03/2019.pdf.

60. Maarten Sierhuis, "Autonomous Systems with Humans-in-the-Loop: The Role of Edge & Cloud Computing," *Stanford Public Seminar Series*, October 17, 2019, https://asia.stanford.edu/wp-content/uploads/2019-10-17-Dasher-Panel-sierhuis.pdf.

61. Mark Wilson, "75% of Ikea's Catalog is Computer Generated Imagery," *Fast Company*, August 29, 2014, https://www.fastcompany.com/3034975/75-of-ikeas-catalog-is-computer-generated-imagery.

62. The author would like to disclose that the next two paragraphs previously appeared in an earlier form in an unpublished grant application submitted to the Hong Kong Research Grants Council.

63. Chi Myŏnghun, "KAIST inmunhak sahoe kwahak e AI chŏmmok hae yunghap yŏn'gu punya kaech'ŏk," *Dong-a Ilbo*, April 8, 2022, https://www.donga.com/news/Society/article/all/20220407/112752744/1.

64. Lee Jisu and Lee Hye-Eun, "Digital Humanities and New Directions in South Korea," *Digital Scholarship in the Humanities* 34, no. 4 (2019): 783, https://doi.org/10.1093/llc/fqy081. I find this survey problematic. For instance, the authors do not explain how they sampled the small number of respondents, and their questions regarding "digital humanities methodology" and "big data analysis" are vague. Nonetheless, I believe this study is sufficient for use as a convenient reference.

65. Lee Jisu and Lee Hye-Eun, "Digital Humanities and New Directions," 784.

66. Darnton, "Blogging, Now and Then."

Acknowledgments

This article is a substantial revision of a piece originally published in Korean in the POSTECH journal *Civilization & Frontier* (*Munmyŏng kwa Kyŏnggye* 문명과경계) in 2020. It is republished here with the kind permission of the editorial boards of both *Civilization & Frontier* and *Korean Studies*. I would like to acknowledge the support of the Innovative and Pioneering Young Researchers Scheme (*ch'angŭi sŏndo sinjin yŏn'guja chiwŏn saŏp* 창의선도 신진연구자 지원사업) at Seoul National University and the Dean's Development Fund in the Faculty of Arts at the University of Hong Kong. Some portions benefited from lab meetings and discussions with Andy Choe, Nabanita Dash, Christina Han, Eugene Jang, Yennie Jun, Kim Juyeon, and Kim Kyoheon. Jerry Wang, an engineer with over 19 years of experience in the cloud industry, clarified a range of questions regarding the operation of data centers and information communication technology infrastructure. Mark Graham, director of the Wayback Machine, responded graciously to my queries. Allison Van Deventer masterfully copyedited the manuscript. I am appreciative of their help and generosity. All errors discovered after the article's publication are my responsibility.

References Cited

Alsop, Thomas. "Server Shipments Worldwide from 2010 to 2020." *Statista*, November 28. 2022. https://www.statista.com/statistics/219596/worldwide-server-shipments-by-vendor/.

Amazon Web Services. "AWS re:Invent 2016: Move Exabyte-Scale Data Sets with AWS Snowmobile." December 1, 2016. https://youtu.be/8vQmTZTq7nw.

Andrae, Anders S.G., and Tomas Edler. “On Global Electricity Usage of Communication Technology: Trends to 2030.” *Challenges* 6, no. 1 (2015): 117–57. https://doi.org/10.3390/challe6010117.

Anthony, Sebastian. “M-Disc is a DVD Made Out of Stone That Lasts 1,000 Years.” *ExtremeTech*, August 10, 2011. https://www.extremetech.com/computing/92286-m-disc-is-a-dvd-made-out-of-stone-that-lasts-1000-years.

Bandaru, Krish, and Kestutis Patiejunas. “Under the Hood: Facebook’s Cold Storage System.” *Facebook Engineering*, May 4, 2015. https://engineering.fb.com/core-data/under-the-hood-facebook-s-cold-storage-system/.

Barrett, Katie. “On Preserving Memory.” *Internet Archive Blogs*, December 18, 2020. http://blog.archive.org/2020/12/18/on-preserving-memory/.

Bernhardt, Darren. “Instagram Obsessed: Can We Vacation Without Posting Every Moment?” *CBC News*, June 29, 2019. https://www.cbc.ca/news/canada/manitoba/instagram-vacation-sharing-social-media-1.5187677.

Bleicher, Ariel, and Katherine Conrad. “We Thought It Was Just a Respiratory Virus: We Were Wrong.” *UCSF Magazine*, Summer 2020. https://magazine.ucsf.edu/we-thought-it-was-just-respiratory-virus.

Byers, Fred R. *Care and Handling of CDs and DVDs*. Washington, DC: Council on Library and Information Resources, National Institute of Standards and Technology, 2003.

Carroll, Rory. “Why Ireland’s Data Centre Boom is Complicating Climate Efforts.” *The Irish Times*, January 6, 2020. https://www.irishtimes.com/business/technology/why-ireland-s-data-centre-boom-is-complicating-climate-efforts-1.4131768.

Cha, Javier. “Pik teit’ŏ wa inmunhak ŭi mirae.” *Munmyŏng kwa kyŏnggye* 3 (2020): 43–77.

Chi Myŏnghun. “KAIST inmunhak sahoe kwahak e AI chŏmmok hae yunghap yŏn’gu punya kaech’ŏk.” *Dong-a Ilbo*, April 8, 2022. https://www.donga.com/news/Society/article/all/20220407/112752744/1.

Ch’oe Hansŭng. “Neibŏ ka mandŭrŏ kanŭn chincha kattŭn kasang tosi: HD maep ŭi haeksim, kojŏngmil 3D chido e taehayŏ.” September 22, 2020. https://blog.naver.com/naver_diary/222096345088.

Clement, J. “Hours of Video Uploaded to YouTube Every Minute as of May 2019.” *Statista*, August 20, 2020. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

Climate Home News. “‘Tsunami of Data’ Could Consume One Fifth of Electricity by 2025.” *Guardian*, December 11, 2017. https://www.theguardian.com/environment/2017/dec/11/tsunami-of-data-could-consume-fifth-global-electricity-by-2025.

Darnton, Robert. “Blogging, Now and Then.” *The New York Review of Books*, March 18, 2010. https://www.nybooks.com/daily/2010/03/18/blogging-now-and-then/.

Gantz, John, and David Reinsel. *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*. Framingham, MA: International Data Corporation, 2012.

Hawkins, Richard. “Use of Big Data in Historical Research.” In *Big Data in the Arts and Humanities: Theory and Practice*, edited by Giovanni Schiuma and Daniela Carlucci, 77–87. Boca Raton, FL: CRC Press, 2018.

Hern, Alex. “Myspace Loses All Content Uploaded Before 2016.” *Guardian*, March 18, 2019. https://www.theguardian.com/technology/2019/mar/18/myspace-loses-all-content-uploaded-before-2016.

International Data Corporation. “IDC’s Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data.” May 8, 2020. https://www.idc.com/getdoc.jsp?containerId=prUS46286020.

Kaplan, Frédéric. “A Map for Big Data Research in the Humanities.” *Frontiers in Digital Humanities* 2, no. 1 (2015). https://doi.org/10.3389/fdigh.2015.00001.

Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge: MIT Press, 2008.

Kirschenbaum, Matthew G. *Bitstreams: The Future of Digital Literary Heritage*. Philadelphia: University of Pennsylvania Press, 2021.

Kodak. 1999. “Permanence and Handling of CDs.” October 13, 1999. http://web.archive.org/web/19991013135327/https://www.kodak.com/global/en/professional/products/storage/pcd/techInfo/permanence.shtml.

Laney, Doug. “3D Data Management: Controlling Data Volume, Velocity, and Variety.” *META Delta*, February 6, 2001. https://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.

Lee Jisu and Lee Hye-Eun. “Digital Humanities and New Directions in South Korea.” *Digital Scholarship in the Humanities* 34, no. 4 (2019): 772–90. https://doi.org/10.1093/llc/fqy081.

Levi, Amalia S. “Humanities ‘Big Data’: Myths, Challenges, and Lessons.” In *2013 IEEE Conference on Big Data*, 33–36. Silicon Valley, CA: IEEE, 2013. https://doi.org/10.1109/BigData.2013.6691667.

Library of Congress. “Update on the Twitter Archive at the Library of Congress.” January 2013. https://www.loc.gov/static/managed-content/uploads/sites/6/2017/02/twitter_report_2013jan.pdf.

Library of Congress. “Update on the Twitter Archive at the Library of Congress.” December 2017. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

Masanet, Eric, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. “Recalibrating Global Data Center Energy-Use Estimates.” *Science* 367, no. 6481 (February 2020): 984–6. https://doi.org/10.1126/science.aba3758.

McGill, Andrew. 2016. “Can Twitter Fit Inside the Library of Congress?” *The Atlantic*, August 4, 2016. https://www.theatlantic.com/technology/archive/2016/08/can-twitter-fit-inside-the-library-of-congress/494339/.

Miller, Rich. “Inside Facebook’s Blu-Ray Cold Storage Data Center.” *Data Center Frontier*, July 1, 2015. https://datacenterfrontier.com/inside-facebooks-blu-ray-cold-storage-data-center/.

Milligan, Ian. *History in the Age of Abundance? How the Web is Transforming Historical Research*. Montreal & Kingston: McGill-Queen’s University Press, 2019.

Naver. “DATA CENTER GAK.” Accessed November 15, 2022. https://datacenter.navercorp.com/green/green-energy.

Naver Business Platform. *Teit’ŏ sent’ŏ Kak*. Seoul: Iro, 2015.

Naver TV. “Teit’ŏ sent’ŏ ‘Kak’ kirok yŏngsang.” September 27, 2013. https://tv.naver.com/v/86767/list/8547.

Notopoulos, Katie. “Flickr Is Deleting Your Photos Soon. Here’s How To Save Them.” *BuzzFeed News*, January 8, 2019. https://www.buzzfeednews.com/article/katienotopoulos/how-to-save-flickr-photos-deleted-download.

O Kilho. “Ssaiwŏldŭ sajin paegŏp.” *Kilho.net*, October 2018. https://kilho.net/archives/various/2190.

OED Online. “Big Data.” Accessed August 1, 2020. https://www.oed.com/view/Entry/18833#eid301162178.

OpenVault. *OpenVault Broadband Industry Report (OVBI): Q1 2020*. Hoboken, NJ: 2020. https://openvault.com/NEW-SITE-OV3/wp-content/uploads/2021/02/Openvault_Q120_DataUsage_FINAL.pdf.

O’Reilly, Michael. “The Unseen Data Conundrum.” *Forbes*, February 3, 2022. https://www.forbes.com/sites/forbestechcouncil/2022/02/03/the-unseen-data-conundrum/.

Raymond, Matt. “How Tweet It Is! Library Acquires Entire Twitter Archive.” Library of Congress Blog, April 14, 2010. https://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/.

Reinsel, David, John Gantz, and John Rydning. *Data Age 2025: The Digitization of the World from Edge to Core*. Framingham, MA: International Data Corporation, 2018. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

Rogoway, Mike. "Facebook Plans Ninth Data Center in Prineville, Says Total Spending There Will Top $2 Billion." *The Oregonian*, June 12, 2020. https://www.oregonlive.com/silicon-forest/2020/06/facebook-plans-ninth-data-center-in-prineville-says-total-spending-there-will-top-2-billion.html.

Rosenthal, David. "Facebook's Warm Storage." *DSHR's Blog*, October 23, 2014. https://blog.dshr.org/2014/10/facebooks-warm-storage.html.

Rosenzweig, Roy. "Scarcity or Abundance? Preserving the Past in a Digital Era." *American Historical Review* 108, no. 3 (2003): 735–62. https://doi.org/10.1086/ahr/108.3.735.

Salter, Jim. "Microsoft's Project Silica Offers Robust Thousand-Year Storage." *Ars Technica*, November 7, 2019. https://arstechnica.com/gadgets/2019/11/microsofts-project-silica-offers-robust-thousand-year-storage/.

Seoul Transport Operation & Information Service. *2019 Sŏul t'ŭkpyŏlsi kyot'ongryang chosa charyo*. Seoul: Seoul Metropolitan Government, 2019. https://news.seoul.go.kr/traffic/files/2020/03/2019.pdf.

Shehabi, Arman, Sarah Smith, Nathaniel Horner, Inês Azevedo, Richard Brown, Jonathan Koomey, Eric Masanet, Dale Sartor, Magnus Herrlin, and William Lintner. *United States Data Center Energy Usage Report*. Berkeley, CA: Lawrence Berkeley National Laboratory, 2016. https://eta-publications.lbl.gov/sites/default/files/lbnl-1005775_v2.pdf.

Sierhuis, Maarten. "Autonomous Systems with Humans-in-the-Loop: The Role of Edge & Cloud Computing." *Stanford Public Seminar Series*, October 17, 2019. https://asia.stanford.edu/wp-content/uploads/2019-10-17-Dasher-Panel-sierhuis.pdf.

Statista Research Department. "Installed Base of Personal Computers (PCs) Worldwide From 2013 to 2019." *Statista*, September 15, 2016. https://www.statista.com/statistics/610271/worldwide-personal-computers-installed-base/.

Statista Research Department. "Number of Smartphone Subscriptions Worldwide From 2016 to 2021, With Forecasts from 2022 to 2027." *Statista*, August 22, 2022. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

Svrcek, Ivan. *Accelerated Life Cycle Comparison of Millenniata Archival DVD*. China Lake, CA: Life Cycle and Environmental Engineering Branch, Naval Air Warfare Center Weapons Division, US Department of Defense, 2009.

Taylor, Kyle, and Laura Silver. "Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally." *Pew Research Center*, February 5, 2019. https://www.

pewresearch.org/global/wp-content/uploads/sites/2/2019/02/Pew-Research-Center_Global-Technology-Use-2018_2019-02-05.pdf.

Tsydenova, Nadezhda. "Russian Telecoms Firms Want Government Compensation for Data Storage Law Costs." *Reuters*, October 29, 2019. https://www.reuters.com/article/us-russia-internet-operators-idUKKBN1X813P.

Unsworth, John. "Scholarly Primitives: What Methods do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" Paper presented at the Symposium on Humanities Computing: Formal Methods, Experimental Practice, King's College, London, May 13, 2000. https://people.brandeis.edu/~unsworth/Kings.5-00/primitives.html.

Vajgel, Peter. "Needle in a Haystack: Efficient Storage of Billions of Photos." *Facebook Engineering*, April 30, 2009. https://engineering.fb.com/core-data/needle-in-a-haystack-efficient-storage-of-billions-of-photos/.

Wiener, Janet, and Nathan Bronson. "Facebook's Top Open Data Problems." October 22, 2014. https://research.facebook.com/blog/2014/10/facebook-s-top-open-data-problems/.

Wilson, Mark. "75% of Ikea's Catalog is Computer Generated Imagery." *Fast Company*, August 29, 2014. https://www.fastcompany.com/3034975/75-of-ikeas-catalog-is-computer-generated-imagery.

Yoo, Tony. "Amazon's New Tool for Startups is a Massive Truck." *Business Insider Australia*, December 1, 2016. https://www.businessinsider.com.au/amazons-new-tool-for-data-hungry-startups-is-a-massive-truck-that-drives-to-the-cloud-2016-12.

Zuboff, Shoshana. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: Public Affairs, 2019.