Guest Editor's Introduction

Digitalization of Korean Studies Materials: An Overview

Kim Byong-sun

Introduction

The purpose of digitalizing Korean studies materials is to reproduce and to compile primary data, text materials, and secondary data including catalogues, indexes, metadata, and others in the fields of humanities and social science studies related to Korea and Korean culture in terms of computer and information technology. It is also to store the reproduced and compiled information in computers so that information can be provided to users in an efficient manner, and to develop computation tools for research. Primary data not only takes the form of text, but also takes the forms of multimedia such as audio and video.

Ideas on how to deal with the data and how to organize them have existed even before the introduction of computers to the fields of humanities and social science studies. However, those ideas were primarily adopted under the premise of editing and printing the information. From the early stage when computers were introduced and used primarily for Korean studies until today, the range of computer applications was limited to only word processing. It was not easy for scholars of Korean studies who were more used to the custom of paper documents to deal with computers.

Therefore when the initial digitalization of Korean studies was embarked upon, namely in the course of shifting from a paper based research system to an electronic research system, it was inevitable that specialists of some fields such as library and information science, computer science, and Korean studies were

The Review of Korean Studies Vol. 8, No. 4 (5-18)

called together to find ways of cooperation. Korea was lagging behind the advanced countries like United States, Japan, and others in hardware, software, and informational technology, so it was reasonable to pursue international collaboration with them.

It was significant that the international conference on the study of digitalization of Korean studies materials was held by The Academy of Korean Studies (AKS) in 1981 in the context mentioned above. I would say the year 1981 was the first year of the era of digitalization of Korean studies. Introducing new technology and an advanced method of scientific research was a necessary process although institutions abroad had performed the leading role of digitalization of Korean studies. HRAF (Human Relations Area Files) of Yale University in the United States, and the National Museum of Ethnology of Osaka in Japan were the foundation for compiling the catalogue of scholars of Korean studies, which was the main theme of the conference. HRAF participated with HABS (HRAF Automated Bibliographic System) in compilation and the Museum of Ethnology not only supported the budget of developing the input/output system of Chinese characters for the catalogue but also carried some projects to actually process and retrieve information (Koh Hesung Chun 1982: 81-103).

All the more, scholars of Korean studies were shocked at the fact that Korean genealogical records were being processed at the Institute of Korean Studies, Harvard-Yenching, outside Korea (Wagner 1982: 111-114). Although both projects were carried out without any suitable input/output systems of Korean or Chinese characters, they still played an important role in stimulating scholars of Korean studies. At that time, Korean scholars of humanities had no idea to apply it to their research and were even ignorant of the computer itself.

Now, a quarter of a century has passed. Many transitions occurred during that time. There was a spread of 8-bit personal computers in the middle of the 1980s, and 16-bit computers were developed in the late 80’s, the so-called IBM compatible computers were welcomed by scholars. Consequently, sophisticated 64-bit computers occupied the desks and laps of scholars. In the mean time, there was big progress in information technology and in designing and producing information processing devices in Korea. Especially, in terms of infrastructure of information and communication, internet and mobile terminals like cellular phones, Korea has secured the highest position in the world these days.

The data processing abilities in the fields of Korean studies has advanced greatly. Character processing has not produced any problems, and great quantities of data can be processed rapidly; obstruction in data transmission is no longer a problem, nor is data storage space a worry. Having been developed from trivial data accumulation of individuals or individual societies, a number of institutions are participating in the construction of a huge humanities database. Large-scale projects of digitalization initiated by the government included the scholastic Korean studies materials, and private companies that are experts in scholastic information were also established.

This paper is devoted to outlining the course of digitalization of Korean studies materials as well as to discussing some basic problems connected with digitalization and its future prospects.

Korean Studies Materials and Processing of Characters and Documents

There are languages and characters in the human world, and the fundamental task of processing Korean studies materials with a computer includes the characters that Koreans have traditionally used. In the primary stage, *hangeul* was used on computers with English operating systems originating in the United States, allowing entry and editing in *hangeul* at the level of word processing. *Hangeul,* with its scientific principle of construction and intellectual system, incorporated well with the information system of computer. However, the processing of some characters such as extended Chinese characters or old *hangeul* essential to classical documents proved very difficult. In addition, software was not designed for Korean characters but for English characters, thus many problems occurred.

Processing of *hangeul* completely at the level of the operating system was obstructed by the observance of the international standard of information exchange (ISO 2022), so efforts in *hangeul* character settlement in individual applications could not be made, especially in word processing, and as a result, several solutions were arrived at. One of it was adapting the foreign word processor to *hangeul* documents in the mid-1980s when the personal computer was placed in the offices of universities and institutions of Korea. The most representative software was “Gem Word Processor” of Trigem Computer Inc. This program offers some kind of modules for data processing as well as document authoring so that textual processing and data retrieving were possible. The WYSIWYG (What You See Is What You Get) supporting the “HWP Word Processor (or Arae Hangeul)” that appeared in late 1980s offered the intelligent processing function of inputting data using various character sets and processing modules, and the sort and macro functions gradually attracted scholars of Korean studies.

The “HWP Word Processor” could handle a combination of a system of *hangeul* and a standard system of Chinese characters (the so-called KSSM method) when it was first announced. Supported by the Korean Computing Society that was concerned with investigation of processing old *hangeul* and the Institute of Korean Historical Materials that was concerned with Chinese characters, the “HWP Word Processor” improved gradually as an intellectual device for processing Korean studies materials (especially input/output) and for multi-language processing, too. The “HWP Word Processor” was a suitable tool for input/output and editing Korean studies materials but was not adequate to organize and store large-scale data. However, there was no such program necessary in database management systems to process and support multi languages within the “HWP Word Processor,” so the method to produce the KWOC (Keyword out of context) index with the “HWP Word Processor” was developed (Kim Byong-sun 1992: 315-392).

The MS Windows adapted Unicode as the operating system of personal computer evolved, and the problem of character processing was solved at the level of the operating system itself in the early 2000s. Therefore, Unicode became effective on personal computers; “HWP Word Processor” abandoned its own code set, adapted the Unicode system, and supported the conversion tool between them. Thus, *hangeul* and Chinese character processing became convenient across range of applications.

Unicode provides a unique number for every character, no matter the platform, program, or language. It represents every character of the world in 16 bits (2 bytes).[1] Many major computer companies like Microsoft, Apple, Sun, HP, etc., and all of the groups which have their own character system go along with the Unicode Consortium. Moreover, in close cooperation with the International Organization for Standardization, Unicode and ISO/IEC 10646 plans are merged in the Unicode system so that it has become the most powerful coding system in the world.

Hangeul and other characters used in the Republic of Korea are enrolled in Unicode in collaboration with some other countries that share the same characters. *Hangeul* is also the official writing system of North Korea and among the

[1] Refer to the Unicode Consortium website: www.unicode.org.

Chinese-Koreans who are one of the ethnic groups in China; thus multi-lateral discussions and co-research are inevitable. However, the Republic of Korea has led the working group at least on *hangeul*, and especially on old *hangeul* characters.

In the Unicode plane, *hangeul* occupies two regions and has registered 11,172 syllabic characters of *hangeul* composed of initial consonant, middle vowel, and final consonant, and one that registers individual initial and final consonants and middle vowels (the so-called *jamo*) that combine to represent a syllable.[2]

> *Hangeul* Syllables: Range: AC00-D7AF (11,172 characters)
> *Hangeul* Jamo: Range: 1100-11FF (240 initial consonants, 90 middle vowels, 66 final consonants, and two fill code characters)

In current computer application programs (for example, the *hangeul* version of MS-Word 2000 and later), modern *hangeul* usually uses completed syllables, and old *hangeul* embody a combination of initial-middle-final *jamo* and the fill code characters in the *hangeul* jamo range of Unicode. Hereafter, there is a movement to handle all *hangeul* syllables with a combination of *jamo*, which most suitable matches the principle behind the invention of *Hunmin jeongeum*.

Chinese characters are the concern of mainly China, Japan, and Korea, and these characters are known as the "CJK Chinese characters" on this account. The fundamental system is generally designed based on the shape of and the composition mode of Chinese characters, which individual countries processed according to their own operating system. In Korea, for example, to input a Chinese character, the user should first press the Chinese character conversion key, and then select one of the characters that are registered as having the same *hangeul* sound. That is the most common method in the Hangeul windows system and *hangeul* applications. Chinese characters are arranged as shown in the following list, and the currently used range is the first part consisting of almost 20,000 characters.

> CJK Unified Ideographs: Range: 4E00-9 FBB (20,924 characters)

[2] MS Word adapts the two kind of representing method that complete system for modern *hangeul* and combining system for old *hangeul* (Shin, Kim, and Ahn 2002: 120-152).

CJK Unified Ideographs Extension A: Range: 3400-4DBF (6,582 characters)
CJK Unified Ideographs Extension B: Range: 20000-2A6DF (42,711 characters)

After arranging all of the Chinese characters in the Unicode range, there are still some areas to study the application and information treatment of Chinese characters at the operating system level and in individual applications. In particular, including the combining system of strokes of Chinese characters, some methods connected with the intellectual processing of Chinese characters have to be researched jointly by Korea, China, and Japan. On the other hand, in saving data and documents, there were a lot of changes. The documents of earlier periods were stored in the peculiar form of the word processor that was the input tool. So, certain problems with compatibility occurred. As the scale of data and the need for standard document form grew, some standard document forms have been developed which are independent of application programs or operating systems or type of computer used. XML is entering into the spotlight and has substantially become the most powerful standard document type.[3]

Born as an extension of SGML (Standard Generalized Markup Language), XML (eXtensible Markup Language) has become the standard document type of electronic texts using hypertext on the internet, and furthermore, it is concerned mainly with displaying the contextual meaning of the documents. Thanks to this advantage, XML is used nowadays for efficient data management. Moreover, it is common to adapt the XML format for browsing data with internet browsers. Today, the public domain of the Korean studies database is stored and represented by this XML form. According to this tendency, some applications such as “HWP Word Processor” and “MS Word” that set out as word processors now support the XML format in its current version, so XML documents can be edited, stored, and relayed to the servers.

In the early 2000s, humanities scholars sought for a way to practically apply XML in the fields of humanities, and as a result, every database of literature is now compiled in XML document structure from the mid-2000s onward. Therefore, for the homogeneity of the humanities database, former documents compiled in the forms of word processors, spreadsheets, or database tables should be converted into the XML format. This cannot be processed in a simple document conversion program; rather, experts accustomed to document content

[3] Refer to the web site of XML Consortium. www.xml.org.

must participate in the designing of the DTD (Document Type Definition) or schema, and the markup of individual documents should be done by them. Various applications for Korean studies materials that support the XML format are expected to develop in future.

Human Resources for Digitalization

The most decisive factor in the digitalization of the Korean studies materials is human resources. Ideal human resources are persons who are qualified both in humanities research and in digitalizing. Otherwise, experts in humanities and in information technology will have to cooperate. As an aside, a person who is competent in information technology does not often have adequate knowledge of the humanities, and in reverse, humanities scholars are often not interested in information technology. For this reason, nurturing professionals covering both areas is an urgent task.

However, in our climate where interdisciplinary educational research is not common, adopting such a methodology in which the scholars of humanities and information technology will cooperative is a difficult issue. Specially, because of low budgets for digitalization projects of Korean studies materials, the chance that information technology experts will participate in these projects is very limited. Therefore, those humanities scholars who are somewhat knowledgeable about information technology are leading the projects, and humanities computing as a science should be developed substantially hereafter.

Scholars who specialize in humanities computing take charge of planning and supervising the projects and researching the field, but there is still a shortage of manpower in charge of driving the projects onward. Thus, lack of suitable manpower for such large-scale digitalization projects is real.

Even in the period of the IMF crisis in Korea, this was the decisive factor in propelling digitalization projects through the public labor system. The quality of the results was not suitable to serve the immediacy of the internet, so the database needed to be reconstructed and corrected extensively; there were also some databases that could no longer be serviced. The ordering institution had no experts in supervising the projects; the executive company had no experience in humanities-related projects, and the laborers (unemployed or not yet employed at that time) did not understand the content itself and were not suitable for the projects. Yet, such a phenomenon occurred.

In the course of time, the digitalization of Korean studies materials has progressed steadily. Such were the content centered companies that were leading the large-scale projects of Korean studies. Specially, starting its missions in developing the font system for Korean studies data processing, Seoul Systems Company secured researchers of its own and accomplished a monumental work through digitalization of the translated edition in *hangeul* of *Veritable Records of the Joseon Dynasty*. Dongbang Media Company Ltd. succeeded Seoul Systems Company's achievements and developed several databases of Korean studies such as the web version of the *Encyclopedia of Korean Culture*, and Nuri Media Inc. published the journals of Korean studies via the internet and concentrated their effort to compile the dictionary of Korean studies. Palman Systems Company devoted itself to the digitalization of Korean Buddhist canons.

Various fonts for Korean studies materials were developed through the participation of these companies, the technology to process Korean studies documents developed, and the research on searching, retrieving, and servicing the information was promoted. These companies announced their products in the form of CD-ROM medium in the beginning, but they have now mostly diverted their content supply to on-line services. On the other hand, developing the contents of Korean studies was achieved in some universities or professional research institutes, and some established organizations to take complete charge of this. The most advanced of these institutions is AKS in this respect. After establishing the Korean Studies Information Center in 1996, AKS had begun to service The Digital Korean Studies database made from the raw materials of the *Encyclopedia of Korean Culture* through the Samsung Unitel Network. Henceforth, some huge nationwide projects of digitalization such as the Digital Library of Korean Studies, the Integrated Historical Information System, and the Electronic Compilations of Korean Local Culture have been propelled forward.

On the one hand, the Institute of Korean Culture at Korea University focuses their research program on the language, literature, and thought of Korea. On the other hand, it runs several laboratories related to dictionary editing and Korean language computing and is intending to research the methodology of the digitalization of Korean studies as well as to compile database construction projects at its Korean studies information center. It compiled the bibliography of translated Korean literature articles and the *Thesaurus of Korean History Terminology*, and now plans a Korean Studies Multimedia Database.

The Language Information Institute at Yonsei University was established in 1986. It is making efforts towards new advanced disciplines such as the grafting of Korean linguistics, lexicography, and corpus-based linguistics onto computing technology. And it offers a retrieval service for the Yonsei Electronic Dictionary of Korean Language and also for the terminology of Korean language computing, such as language and linguistics, information and communication, computer science, and cognition science, etc.

The Humanities Information Institute at Seoul National University, a branch of the Humanities Research Institute, was established in 1997. It constructs the system of information related to the humanities area to advance the humanities research capacity in accordance with developments in computer science. So far, its activity does not go beyond publishing the "Korean Modern Literature, 100 Years, CD-ROM" in 1998.

Although the institutions on Korean studies did originally set out to digitalize information, they have been promoting their information capacity by taking part in government-supported Korean studies information projects since the late 1990s. Representative institutions participating in the Integrated History Information System include the Gyujangak Archives at Seoul National University, the National Institute of Korean History, and the Society of National Culture Propulsion, etc. These institutions have extended their original functions, such as editing, compiling, translating, and researching Korean studies raw materials not by establishing large-scale organizations like information centers but by maintaining departments of data information or computing teams. Specially, as internet services are progressing, the function of information processing and its public service occupies the greater part of the institutes' activities.

Educating potential manpower in humanities computing is uncommon at universities. At present, several Korean language departments are interested in humanities computing. The Graduate School of Yonsei University established a co-curriculum of Korean language computing that is run by the departments of linguistics and computer science, and some departments of Korean language and literature offer one or more lectures on language computing in their ordinary curriculum. However, there is no department majoring in the humanities and linguistic computing so far. Similarly, in the case of the Graduate School of Korean Studies at AKS, computation of Korean studies materials is included as a basic subject, and some advanced subjects on computing related to individual majors are listed in their curriculum as well. Efforts to make humanities computing an independent major in the near future are being continued.

In addition to these, some universities run lectures on cultural content, reorganize existing departments, and employ professors in this field. One of them is the Graduate School of Cultural Technology at the Korean Advanced Institute of Science and Technology (KAIST). However, such institutes are mostly interested in the digitalization of contemporary popular art, and they put emphasis on audio-visual materials of Korean culture. Consequently, specially trained experts on humanities computing who have knowledge on both the humanities and technology of information science have not yet appeared.

All the institutions and companies doing humanities computing projects are experiencing difficulties with data input, correction, and special processing of data, and such difficulties are due to lack of qualified manpower. In the mean time, the Korean Studies Information Center at AKS has contributed in educating the people in humanities computing by carrying out certain huge projects on Korean studies digitalization. It transformed the former department of computing and information to that of humanities computing this year so that it now develops computer programs and technologies on Korean studies materials. Also, it is expected that the Center will install laboratories and run some educational programs whereby graduate students can master information technology that is specialized towards Korean studies materials.

Future Prospects

As observed in former sections, the digitalization side of Korean studies materials has increased during the last ten years. The most active part has been the input of text data, and there should be continued investment and interest for images and multimedia data, which are fundamental to Korean studies. And those projects that not only make new contents but also investigate cultural phenomena such as the project of the Electronic Compilation of Korean Local Culture should be pushed ahead. At the same time, formerly compiled electronic data has to be reformed and adjusted to meet new information technology standards. Besides such efforts of collecting and compiling, activities and business related to the ones mentioned below also need to be invigorated.

Standardization of data and systems is the most urgent and important task at hand. In fact, the effort for standardization of Korean studies materials computing is as yet lacking. In the process of XML documents, the enactment of DTD or schema, and the establishment of markup tags are different in participating institutions and in each project. There are inconsistencies in the data compiled in different times. Because of the rapidly developing information environment, the older data format has become out-of-date. Therefore, this standardization effort should be continued for a while. In addition, the initiatives that manage the standardization of humanities computing should be established to improve the efficiency and consistency of the projects of digitalization on Korean studies materials.

As the projects continues, the accumulated data amounts to thousands of gigabytes, so studies on the method of information acquisition with such a large-scale database should be carried out. The present stage needs data to be well-organized in such a way that intellectual processing can retrieving it, whereas former efforts concentrated mainly on gathering the data itself. Statistical approaches to data have to be developed theoretically and practically in the humanities area. Compared to the social sciences area, it is not common for scholars in the humanities to apply statistic quantitative data to their research. It is also a task of humanities to train experts in this field.

There are so many associated fields connected to the digitalization of Korean studies materials; humanities scholars should thus become sensitive to the trend of informational technology. Generally, not only the development of computer science and technology of communication but also the technology of Korean natural language processing should be kept in mind. This is because most Korean studies materials have been compiled in the form of text, and this text data is linked closely with natural language processing. As far as the natural language processing is developed, the automatic keyword extraction or automatic summarizing techniques could be advanced further. As well as scholars of Korean linguistics, scholars of computer science are also interested in natural language processing. So, cooperative research is the most desired mode for the field to move forward, and the attendant results should contribute greatly to the data processing of Korean studies.

On the other hand, scholarly exchanges with the North Korea may increase during the 21st century and on. North Korean scholars in this field have been interested in language information processing technology since the early 60s, and actually started the development of an automatic translation machine from Korean to Russian and Korean to Chinese (Kim Byong-sun 2000). By activating exchange between North and South Korea, the digitalization of Korean studies raw materials that North Korea retains, the employment of North Korean experts to input data, correction and compiling, and the sharing and utilizing the common data or technology of information science will become possible.

We need to consider the mind of scholars of Korean studies who take advantage of the database but have not yet adapted to informational society and remain in the realm of paper-based culture. Although much data have been accumulated, they are not being utilized well by scholars of Korean studies. Thanks to computer communication, the intellectual environment around us has been transformed to that of information and knowledge. This should enable scholars to escape from their attitude and method of relying on paper book. Korean studies scholars can now access the database, collect necessary data, process it in a desired format, reorganize it as intellectual form, and interpret it comprehensively. A powerful synergy phenomenon will result if information technology with the infrastructure of internet communication and Korean studies materials are merged together.

Now, some individual essays will inspect the present condition of digitalization of the field of Korean studies and introduce the content and progress of important projects in detail.

References

Bak Jae-su. 1999. *A Study on Linguistics of DPRK.* Pyeongyang: The Academy of Social Science.

Kim Byong-sun. 1992. *Korean Language and Computer.* Seoul: Hansil Publications.

______. 1994. “Fundamentals in Data Processing of Korean Studies Materials.” *The Review of Korean Studies,* Vol. 56. Seongnam: The Academy of Korean Studies.

______. 1998. “Hangeul Word Processor; Current Tasks in Challenge and Response.” *Loving Hangeul,* Vol. 8. Seoul: Hangeul-sa.

______. 2000. “The Present State of North Korean Language Information Processing Studies.” *The Journal of Korean Studies,* Vol. 2. Seoul: Seoul Branch of the International Society for Korean Studies.

Koh Hesung Chun. 1982. “The Computerization of Bibliographical Information and the Directory of Researchers in Korean Studies.” *The Proceedings of the Conference on the Computerization of Korean Studies Materials.* Seongnam: The Academy of Korean Studies.

Korean Computing Society. 1998. *The Journal of Korean Computing,* Vol. 2. Seoul: Korean Computing Society.

Shin, Kim, and Ahn. 2002. “Representation of Old *Hangeul* at MS Word 2000.” *Compilation of Research Materials of Korean Language 1*. Seoul: The National Institute of the Korean Language.

Wagner, Edward W. 1982. “Problems in the Computerization of Materials in the Korean Studies Field: A Report on the Munkwa Project.” In *The Proceedings of the Conference on the Computerization of Korean Studies Materials.* Seongnam: The Academy of Korean Studies.

www.koreandb.net.
www.unicode.org.
www.xml.org.