On-line full text resources on Korean Studies from the Google Library Project

Yunah Sung

Asia Library, University of Michigan

Abstract

This paper presents an overview of the Google Book Search and Google Library Project, and discusses the specifics and implications of this massive digitization of library collections for studying and research on the field of Korean studies.

The Google Library Project was initiated in December 2004 in an attempt to build a global digital collection of books from Harvard, the University of Michigan, New York Public Library, Oxford and Stanford University libraries, and it has expanded the partnership to 28 libraries in 8 countries.

These digitized materials will be fully searchable, allowing users to employ key words to search the indexes, tables of contents and full text of books. The full text of public domain materials can be viewed and printed from the Google site. Books that are still under copyright will show up in search results, but users will only be able to view very small text snippets and/or bibliographic information. Also "Find this book in a library" function takes a user to the OCLC WorldCat, where the user can find a local library that has the physical book.

This project will make it possible for researchers to access the full text of primary western language resources in the public domain on Korean studies over the Internet without any restrictions. Also any users who are not familiar with the ALA-LC Korean romanization rules, will be able to locate an item in the Google Book Search site by searching a keyword in Korean. The biggest hurdle for digitized resources in non-roman characters, particularly in Korean Han'gul and Hancha, is the OCR process. Materials in Korean are more prone to OCR errors and thus the ability to retrieve every occurrence of a search term will be diminished. It is highly expected that these software related technical difficulties will be solved in the near future.

This is an on-going project and more full-text materials will be globally available as Google massively digitizes print collections held by partner libraries.

Introduction

Google, the most used search engine on the web, has been working to create a global book repository in an attempt to achieve its mission which is to "organize the world's information and make it universally accessible and useful." The domain google.com receives several hundred million queries each day and average of 125 million users per month visits the site in 2007-2008. [1]

Google Book Search, one of Google's products such as Google Maps, Google Scholar, and Google Earth, was developed to create a virtual book repository for fully searchable full-text books in all languages that helps users locate relevant books more easily and effectively. When users make a query in Google, they first see a list of relevant websites

by default. When they are interested in an image that matches the query, they click on "Images". In the same way, users can see a book that matches the query in Google's search results by clicking "Books". Once taken to "Books", that is the same as "Google Book Search" site, users can see a web page displaying a scanned image of the relevant page from the book, as well as multiple links to online book stores for purchasing and another link to local libraries for borrowing physical copy of the book. In order to provide this service, Google needed full-text of all these books in its depository, which primarily come from two different sources; Google Partner Program and Google Library Project.

Google Partner Program is for publishers and authors. Publishers and individual authors who have joined the Program send a digital copy of a book in PDF electronically to Google or ship a hard copy of the book to Google to be digitized. The benefits to participating in Google Book Partner Program are to promote the book for free on Google, to increase book sales at online or offline bookstores, and to reach out to targeted users easily. For further information on Google Book Partner Program, visit its site at https://books.google.com/partner/

In other hand, Google Library Project is for partner libraries to digitize the print collection of each library and include the digitized collections in Google Book Search. The Google Library Project was initiated in December 2004 to digitize several millions of books held by major research libraries in US and the UK, including Harvard, Michigan, New York Public, Oxford, and Stanford. Google is now working with about 28 libraries in US and in the world, including Bavarian State Library in Germany, Complutense University of Madrid in Spain, Ghent University Library in Belgium, Keio University Library in Japan, Lyon Municipal Library in France, National Library of Catalonia in Spain, and University Library of Lausanne in Switzerland. Its website (http://books.google.com/googlebooks/partners.html) provides updated list of partners.

Google Library Project at University of Michigan

The University of Michigan Library has been taking a leading role in digitally preserving scholarly research materials from the Library's important collections since the mid-1990s. In an effort to continue the goals of digitization projects, the Library has partnered with Google to digitize the entire print collection of the Library and to offer those books available for searching in the library catalog, Mirlyn, as well as in the Google Book Search. In return, the Library receives digital files of these works as preservation copies from Google. Even though some people were skeptical about the project in the beginning, the Library was convinced that the Google Library Project will create new ways for users to search and access the Library content, opening up the Library collections to users at the University of Michigan, as well as users throughout the world. [2] Also the Library believed that this kind of massive digitization project will provide great opportunities to the Library to completely digitize the entire collection.

President Coleman strongly supported this ground breaking initiatives and discussed the importance of the Google Library Project at the annual meeting of the Association of American Publishers in February 2006. She pointed out that; it will protect thousands of important publications from disappearing due to damage and decay, while making their contents searchable by anyone with an Internet connection. "The University of Michigan's partnership with Google offers three overarching qualities that help fulfill our

mission: the preservation of books; worldwide access to information; and, most importantly, the public good of the diffusion of knowledge," she said. [3]

While some of the Google Library Project partner libraries limit their contributions to public domain books only, The University of Michigan and Stanford University have decided to digitize the entire print collection of the library. The reasons behind this decision are that the paper quality of the collection and the term of the "public domain works". Regarding the paper quality of the collection, it was estimated that more than 1.5 million volumes of the Library collection are brittle and another 3.5 million books on acidic paper are at risk. Once digitized, these materials can be conserved and users can access to digital copies. In an effort to preserve these collections, the Library had been digitizing about 5,000 – 8,000 volumes annually. [4] In regard to copyright status, it is believed that every book in the Library will eventually fall into the public domain and will be fully accessed by the pubic. In February 2008, The Library announced that the Library has put the millionth book on-line, out of the 7.5 million volumes in the library's current holdings. Most of these on-line digitized collections are products of the Google Library Project which digitized an average of 20,000 volumes a week. [5]

MBooks, these digitized collection, can be searchable in the Library online catalog, Mirlyn. In compliance with copyright laws, full-text of works that are in the public domain or out of copyright are available. Mirlyn also carries links to the Google versions of the work.

As one of the leading research libraries in US and in the world, the University of Michigan Library has been committed to provide the users on / off campus with a comprehensive library collections and keep them preserved and accessible.

Detailed search guides in Google Book Search:

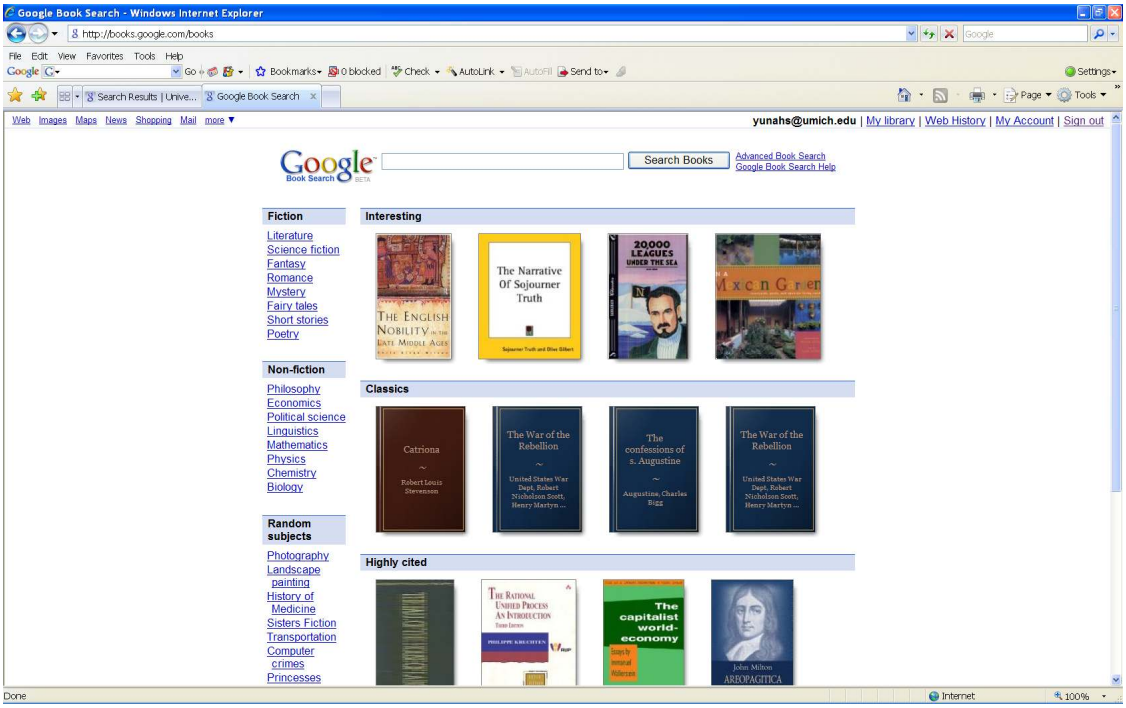

Figure 1. Google Book Search homepage

1. Search options

There are two search options at Google Book Search; default search and “Advanced Book Search”. For the default search, users enter a query into a search box and click the “Search Books” button, just like any other Google searches. For more specific searches, users can go to the “Advanced Book Search” located next to the “Search Books” button. “Advanced Book Search” page provides several search boxes and options to limit the search results. Some of these options are language of the book, publication date, and ISBN. When Google finds a book with content that contains a match for the users’ query, the book is linked to in search results.

During Google search process, Google finds a book by searching for the specific query throughout the entire text of the book, as well as bibliographic information described on its cataloging record supplied by OCLC (Online Computer Library Center). If a query term is not specific enough, Google could retrieve more than several thousands books. While it is not effective for a user to review all these books, the user has much higher possibility to discover the book which couldn’t be located by searching the term only from bibliographic information.

2. View options

There are four different types of views, depending on the book’s copyright status and publisher/author wishes;

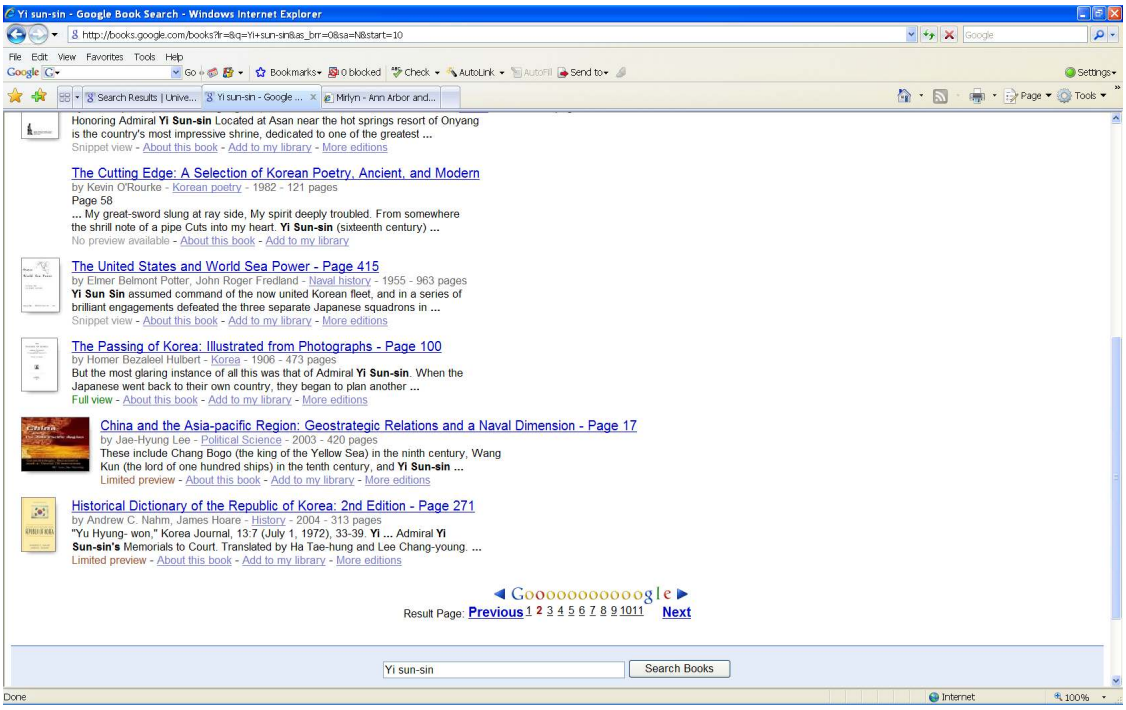

Figure 2. View options

Full view:

If a book is out of copyright, or if the publisher or author has given Google permission, a user can see entire books in Full View online.

Limited preview:

If the publisher or author has given Google permission, a user can see a limited number of pages from the book as a preview.

Snippet view:

If a book is under copyright and the publisher or author is not part of the Partner Program, a user can see bibliographic information about the book and a few snippets – a few lines of text to display a user's search term in context.

No preview available:

For books where Google is unable to show snippets or previews, a user will see only basic bibliographic information about the book.

3. Detailed information about the book

"About this book" tab, located on the top pane, includes in-depth bibliographic information, such as title, author, publication date, pagination, and subjects, as well as a source of digitized library, if available. Users may also find additional information, such as cover of the book, key terms and phrases, chapter titles, a list of related books and books to use as references. Every book has links directing to bookstores where a user can buy the book and to libraries where a user can borrow it.

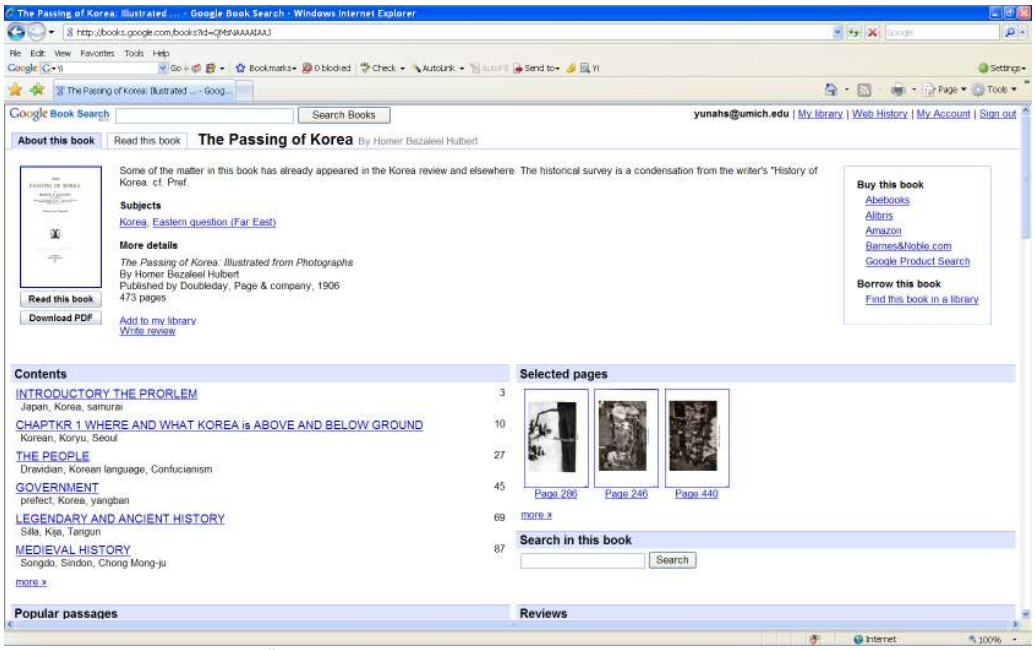

Figure 3. "About the book"

Some books contain a short review of the book supplied by a publisher or an online book seller. Also some books include “Subjects” which are mostly Library Congress (LC) subject headings, but some books also include non-LC subject headings.

A series of key terms and phrases are automatically populated from the text of the book. In many cases, each key term is hyper-linked to “Search in this book” feature and Google shows each page of the book the key term appears.

4. Search on entire of the book

One of the most useful features of the Google Book Search for studying and research is “Search in this book” in all digitized materials. Whether a book is in the public domain or in copyright, users can search a word or phrases within the book by entering the term in the “Search in this book” box. Google will show how many times the term appears on the book and display a list of page numbers where the term appears. Instead of reading throughout the entire book, users can easily pinpoint certain pages of the book where the term is mentioned.

5. Preview or read the book

For those books still in copyright, users can preview limited number of pages of these books, which are displayed with the permission of publishers and authors. Users only can see certain pages which the copyright holder has made available.

For those materials in the public domain, users can view full-text copies for free.

6. Download or print-out public domain works

If a book is out of copyright (generally works before 1923), or if the publisher has opted to allow, users can download, save, and print the entire book in PDF files for free. These downloaded PDF files are image files without keyword search capability. Google placed technical restrictions on automated querying on PDF files to prevent any abuse by commercial parties. Users should read the book while connected to Google Book Search in order to take full advantage of hyperlinked table of contents and highlighted keywords.

7. Google Maps in Google Book Search

For some books, Google Book Search is integrated with Google Maps. If Google detects names of specific locations in a book, a tagged map with page references will be displayed on the “Places mentioned in this book”. When a user clicks on a place name, Google Map displays the location on the map as well as page numbers in the book where the place name appears. Also Google displays the same information when a user directly clicks on a place on the map. [6]

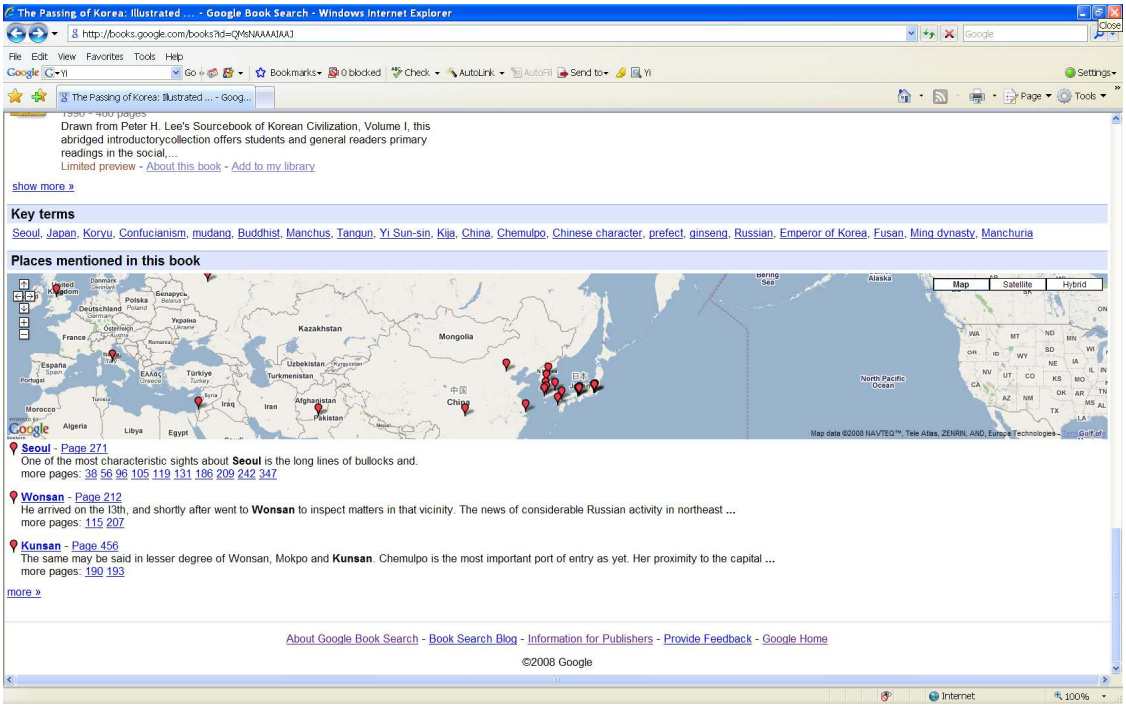

Figure 4. “Places mentioned in the book” on Google Maps

8. Add to “My Library”

For a Google account holder, this book information can be added to “My Library” feature. Users can create personalized library collections for selected books by utilizing this feature. These library collections are accessible anytime and anywhere users can log in to the Google account. To build up their own collections, users can click the “Add to my library” link from search results. Also they can annotate it and share it with other people by sending them a link to “My Library”. Also users can set up RSS feeds to alert any updates.

As an example, “Korean Studies” library has been created in Google Book Search. Anyone, even who does not have a Google account, can view this collection at following link; http://books.google.com/books?uid=10027030872909235172 .

Further information on “My Library” features can be found at My Library FAQ site (http://books.google.com/support/bin/answer.py?answer=75375&src=top5&lev=).

9. Access to the print book

If a user would like to locate a printed copy of the book, one has an option to “Buy this book” from selected publishers and booksellers or to “Find this book in a library” from a list of libraries where one can borrow the book.

When users click a link to any bookseller, it directs to the bookseller’s site and automatically searches the availability and price of the book. It is convenient for users to purchase the book online from selected booksellers, but they don’t have to.

“Find this book in a library” feature connects to a selection of entries from OCLC’s Open WorldCat (http://www.worldcat.org/) and displays a list of rending libraries sorted by the proximity of user’s postal code. A user can check the availability of the book by clicking a link to a nearby library.

More detailed explanation on certain features can be found at Using Google Book Search site; http://books.google.com/support/bin/topic.py?topic=9259.

Issues requiring further research and development

1. Refining search results

Even though Google Book Search provides “Advanced Book Search” option, the scale of matching resources could be too overwhelming for a user. There is no sorting option, such as by author, title, date descending or date ascending, within search results.

Google’s algorithm uses PageRank for determining a page’s raking in the search results. PageRank is a numeric value from 0 to 10 that represents how important a page is on the web. This value is critical for a user due to the fact that it is one of the factors that determine a page’s raking in the search results. The higher the number is, the more important the web page is. Those pages receiving the highest number are returned at the top of the user’s search results page with the other results displayed in the order of the rank.

Once Google returns search results based on value of the PageRank, users have no control over rearranging or sorting the search results to further narrow down the result set. They should patiently page through the search results until they find the ones that they are interested in. A query in Google Book Search could return thousands of results and there could be nothing pertinent to the query in the first few screens of results. It is critical to pinpoint a relevant book from a large set of results.

2. User interface

Google is well known for simply and clean look on its homepage at Google.com. But Google Book Search is not shown on Google’s homepage and users should click “More” to get to its site. To make visible to Google website visitors, it is highly recommended that Google makes the link to the Google Book Search directly from its home page.

Navigating within Google Book Search sites is a bit cumbersome. To return to the search result page, a user needs to click the recent pages button on the web browser.

Also, to locate a book efficiently and effectively at Google Book Search, a user has to have very sophisticated search strategy. Google does not provide detailed user services nor assist to meet a user’s needs during the search process. Users could be easily stranded in this massive online collection while reviewing numerous entries that Google retrieved.

It is highly desired that Google would work with publisher partners as well as library partners to provide users with more user-friendly services for all these online collections.

3. Non-Roman languages

There are two major problems found in Google Book Search for Korean language materials; OCR process and romanization.

Google Book Search is using the underlying OCR (Optical Character Recognition) text analysis to create an index to digitized works for keyword searches. In other words, this OCR process is the process of creating machine editable files from paper documents. Owing to this process, users can get the information on which pages the search terms appear and how often the terms appear on each page. While typical accuracy rates for the recognition of typewritten text in Latin-script exceed 99%, OCR software for East Asian scripts, in special Korean Han'gul, is far from perfect.

Google is challenged by OCR problems, which are resulted from the multitude of languages, scripts, and fonts in contributing partners' collections, as well as from the quality of the page itself. Most of times lower quality OCR is resulted from lower quality of the page that is deteriorated and discolored. Currently Google is reprocessing texts and will send much improved versions to contributing partners. [7]

In addition to the accuracy rates, there are more complex problems in OCR. For examples, multiple languages in the same book, texts written from right to left in vertical lines running from right to left (Arabic and Hebrew), and texts written from right to left in vertical lines running from top to bottom (frequently found in pre-modern East Asian materials). It is crucial to improve the OCR which Google produces.

In terms of romanization, diacritics in Korean catalog records have direct impact to search results in Google Book Search. There are three diacritics found in Korean records; alif, ayn, and breve. In contrast to WorldCat searching, the presence of correct diacritic matters in Google Book Search. Here are sample search results;

Characters / Diacritics	No. of books found	Search term
Han'gul	578	대원군
Hancha	883	大院君
Valid diacritics	432	Taewŏn'gun
One valid diacritic	375	Taewŏngun.
No diacritics	371	Taewongun
Invalid diacritics from keyboard	399	Taewon'gun (apostrophe)

Characters / Diacritics	No. of books found	Search term
Han'gul	770	박정희
Hancha	615	朴正熙
Valid diacritics	381	Pak Chŏng-hŭi
One valid diacritic	18	Pak Chong-hŭi (박종희, not 박정희)
No diacritics	568	Pak Chong-hui
W/O hyphen	932	Pak Chŏng hŭi

A cataloging record for a Korean material has been created in OCLC WorldCat in accordance with the ALA-LC Romanization tables: Transliteration schemes for Non-Roman scripts, which is approved by the Library of Congress and the American Library Association. Almost all Korean cataloging records contain at least one diacritic and most of local library online catalogs and WorldCat normalize all diacritics in a search term. However, that's not the case in Google Book Search. It is recommended to Google to further research on this matter so that users do not have to key in valid diacritics.

Concerning romanization and word division in Korean cataloging records, users can receive relatively reliable search results by entering key terms in Han'gul or Hancha as shown above. It is very welcoming news for users, not only in US, but also in all around the world.

Since punctuations and diacritics have direct impact to search results, users are advised to perform several searches with possible variations, in special, personal name in English.

No. of books found	Search term
733	Yi Sun-sin
159	Yi Sunsin
51	Lee Sunsin

4. Google's future

At present, Google is the most heavily used internet search engine in US. However, it is not possible to speculate on future development of Google. As new and powerful search engine equipped with better features and faster retrieval will attract internet users some

time in the future, the number of site visitors at Google will gradually decrease. Even though individual partner library will be able to manage its own digitized collections, Google's long-term plan to manage these digitized resources is not clearly laid out.

Conclusion

The invention of the computer and the advent of the Internet brought about the total transformation of the information landscape. We all spend a great deal of time in front of a computer, anticipating that the computer electronically delivers relevant information at our fingertips, in just one click of the mouse. In accordance with this transformation, the world of the scholarly communication also has been changed very fast in a big scale and academic research libraries around the world have been developing a long and short term strategies to meet users' demands and to facilitate scholarship.

The process of moving physical collections of the library to electronic formats can't be successfully achieved without strong technical and financial supports to the library. In this regard, Google's initiative to massively digitize research materials from selected libraries around the world needs to be highly evaluated. Of course, there are foreseen and unforeseen obstacles laid before all of us. However, these obstacles could be resolved by the worldwide efforts among partner libraries and publishers.

Academic research libraries will continue to manage and provide access to both physical and digital information resources through the application of the latest information technology. The ultimate mission of the library is to support research, teaching, and learning for faculty and students on campus as well as off campus. In an increasingly electronic age, the information needs of the 21st century will be largely supplied across the global digital networks. Google partner libraries need to make the collaborative efforts to collectively manage digitized collections and to seamlessly supply to users in the world in the age of Google and beyond.

Notes

[1] Unique visitors, http://siteanalytics.compete.com/google.com/?metric=uv (accessed August 2008)

[2] U-M University Library-Google Digitization Project: Project Overview (December 13, 2004), http://www.lib.umich.edu/mdp/overview.pdf (accessed August 2008)

[3] Mary Sue Coleman, "Google, the Khmer Rouge, and the Public good" (address to the Professional/Scholarly Publishing Division of the Association of American Publisher, Washington, D.C., Feb. 6, 2006), www.umich.edu/pres/speeches/060206google.html (accessed August 2008)

[4] M Library achievement (December 2007), http://www.lib.umich.edu/about/libraryaccomplishments.pdf (accessed August 2008)

[5] Library news, February 2008, http://www.lib.umich.edu/news/millionth.html (accessed August 2008)

[6] Using Google Book Search, http://books.google.com/support/bin/topic.py?topic=9259 (accessed August 2008)

[7] Languages in MBooks, Blog for Library Technology (August 1, 2008), http://mblog.lib.umich.edu/blt/ (accessed August 2008)

Bibliography

Carr, Reg, 1946- The academic research library in a decade of change / Oxford : Chandos, 2007.

Casey, Michael E., 1967- Library 2.0 : a guide to participatory library service / Medford, N.J. : Information Today, c2007.

Gordon, Rachel Singer. Information tomorrow : reflections on technology and the future of public and academic libraries / Medford, N.J. : Information Today, c2007.

Libraries and Google / New York : Haworth Information Press, c2005.

Miller, Michael, 1958- Googlepedia : the ultimate Google resource / Indianapolis, Ind. : Que/Sams ; London : Pearson Education [distributor], c2007.

Theorizing digital cultural heritage : a critical discourse / Cambridge, Mass. : MIT Press, c2007.