Journal of East Asian Libraries

Journal of the Council on East Asian Libraries
No. 165, October 2017

CONTENTS

From the President 3

Essay
A Tribute to John Yung-Hsiang Lai 4
Eugene W. Wu

Peer-Review Articles
An Overview of Predatory Journal Publishing in Asia 8
Jingfeng Xia, Yue Li, and Ping Situ

Current Situation and Challenges of
Building a Japanese LGBTQ Ephemera Collection at Yale
Haruko Nakamura, Yoshie Yanagihara, and Tetsuyuki Shida 19

Using Data Visualization to Examine Translated Korean Literature
Hyokyoung Yi and Kyung Eun (Alex) Hur 36

Managing Changes in Collection Development
Xiaohong Chen 45

Korean Römaniz’atiön: Is It Finally Time for the Library of Congress to Stop
Promoting Mccune-Reischauer and Adopt the Revised Romanization Scheme? 57
Chris Doll

Reports
Building a “One-hour Library Circle” in China’s Pearl River Delta Region 85
Paul W. T. Poon

Interview with the Curator of the Po Leung Kuk Museum 87
Patrick Lo and Dickson Chiu
Project Report:
  Web-Archiving Chinese Social Media: Final Project Report  93
  Collecting Social Media Data from the Sina Weibo Api  113

New Appointments  136

Book Review  137

  Heng Ge, NYU Shanghai Library
From the President

With my usual gratitude, I reviewed the Table of Contents of the coming JEAL issue. I saw the tribute to one of our pioneering librarians at Harvard-Yenching Library written by our legendary former CEAL president Eugene Wu (1976-1979), providing an invaluable biographical record of a retired colleague; a group of peer-reviewed articles, which showcase the results of the high quality research done by our colleagues in a variety of subjects, including an overview of “Predatory Publishing in Asia,” LGBTQ Collection in Japanese, Data Visualization for translated Korean Literature, questions on the obsolete usage of McCune-Reischauer as the Korean Romanization Scheme by LC; and the project reports that outlined the web-archiving projects supported by the CEAL/Mellon Innovation Grant Program, which represent the newest trend of our profession; and, finally, a book review of the work by a CEAL member on a part of history that has been intentionally blurred in its original homeland.

All of these remind me of the current hot topic: freedom of speech and press. When I saw the attack on the media by our president, who called the media “the enemy of American people,” I had an ominous feeling about the fundamental values of our democratic system and our profession. Are the freedoms of speech and press protected by the First Amendment of our Constitution in peril? As a librarian, I always believe in “letting a hundred flowers blossom” and keeping culture, civilization, and thoughts alive without bias. There is an old Chinese saying that “a thousand-mile journey starts under your feet.” JEAL is our forum for freedom of speech and press; let’s all try our best to water and nurture this flower and keep it blossoming.

Last week, I attended a memorial service for Prof. Wm. Theodore de Bary (1919-2017), former Provost, noted educator, Sinologist, and alumnus of Columbia University, during which I saw the tribute written by his student Prof. Carol Gluck, who cited his 1970 AAS Presidential Speech titled “Nonpolitical but not Unconcerned.” This, after almost 48 years, I found amazingly fits our profession, specifically in today’s climate.

Jim Cheng
CEAL President
A Tribute to John Yung-Hsiang Lai

Eugene W. Wu

John Lai and I met in 1960 in Taipei on my first trip to Taiwan. He was then the head of reader service at the National Taiwan University Library, and I was in Taipei for the Hoover Institution to microfilm the “Chen Cheng Collection,” an archive of some 1,500 documents of the Kiangsi Soviet Republic, 1931-1934. He had just returned to Taiwan the year before from the United States where he earned his MA degree in Library Science at the Peabody Library School, and was involved at that time in planning a library science department for the National Taiwan University. After the Department was established the following year, he was appointed assistant professor at the new department, then promoted to full professor in 1964, and became its chair in 1965. He held that position until 1972 when accepting an invitation from Harvard University to become the head of the newly established Cataloging Department of the Harvard-Yenching Library. He began the second phase of his career in the United States. He was appointed Associate Director of the Harvard-Yenching Library in 1978 and served in that capacity until his retirement in 1996. He served with great distinction at the Harvard-Yenching Library for a total of twenty-four years.

Mr. Lai, as we affectionately and respectfully called him, stands tall in the East Asian library community. He was born and raised in Taiwan and, like many of contemporaries, finished his college education in Japan, where he attended the Tokyo Imperial University. Prior to coming to Harvard, he was already an established library educator and an authority in cataloging and classification. He was best known for the 賴永祥中國圖書分類法 (based on the work by Liu Kuo-chun 劉國均 decades earlier), the standard classification system used by libraries in Taiwan and widely adopted for use by libraries in Hong Kong, Macau and Singapore. The classification system went through eight revisions by Mr. Lai, from 1964 to 2002, after which he generously donated the copyright to the National Library in Taipei. The National Library published a further revision in 2007 under a changed title: 中文圖書分類法, and the classification system remains the system of choice today at libraries in Taiwan. His appointment at the Harvard-Yenching Library was therefore a perfect fit.

For decades prior to the establishment of the Cataloging Department, cataloging in the Harvard-Yenching Library was carried out separately under the direction of the heads of the language sections, using the 漢和圖書分類法, commonly known as the Harvard-Yenching Classification System, devised by Dr. A. Kaiming Chiu 裘開明, the founding librarian of Harvard-Yenching Library. This arrangement worked well for many years, but with the expansion of the Library’s collection scope, resulting in the ever increasing volume and
variety of books and non-book material that were coming to the Library, a decision was made to bring all cataloging work into one independent office in order to provide more unity in practice and also to relieve the language section heads of their supervisory responsibilities over cataloging, so they could devote full time to collection development. But finding someone to head up this new department was much harder than making the decision. Having been a cataloger myself for almost a decade at the Hoover Institution, I knew it would require a person not only of superior technical skills, but also managerial ability, a good grounding in East Asian civilization, fluency in the major East Asian languages, as well as ability to communicate and inspire. It was a worldwide search, and we found in Mr. Lai the perfect candidate. However we were not at all sure he would be interested, for he was then at the summit of the field in Taiwan and leading a good and comfortable life there. So I made a special trip to Taiwan to sound him out. I briefed him on our plan for the new Cataloging Department and told him we would be honored to have him join us at this crucial moment in the development of the Harvard-Yenching Library. He was courteous but noncommittal, saying he would think about it, consult his family and then let me know. It was sometime later that we received the good news. I and my colleagues at the Harvard-Yenching Library were absolutely elated, and we made plans to welcome him and Mrs. Lai to Cambridge. They arrived in Boston on a bitterly cold day toward the end of January in 1972, and Mr. Lai began his tenure at the Library the following month.

With energy and a knack for organization, he hit the ground running. Being cordial and approachable and with a strong work ethic, he quickly earned the support and respect not only from those working under him, but also from his other colleagues in the Library. The Cataloging Department got off to a flying start and ushered in an important phase in the development of the Harvard-Yenching Library.

One of the first tasks he faced was the revision of the Harvard-Yenching Classification System. The system was quite adequate for pre-modern East Asian publications, but somewhat deficient for the modern period. Mr. Lai undertook to do the revision and completed the work shortly, bringing to bear his experience of revising his own classification scheme. These revisions were shared with other East Asian libraries using the Harvard Classification System at that time. In addition to his supervisory responsibilities over cataloging, Mr. Lai also took up the task of cataloging the Library's large collection of Bibles and Protestant missionaries' writings, and a catalog of this collection, comprising 786 items, was published in 1980 by G. K. Hall in Boston under the title: *Catalog of Protestant Missionaries’ Works in Chinese*. Two years later, in 1982, a microfiche edition in 1,750 sheets comprising the content of the entire collection was issued by Inter Documentation Center (IDC) in Leiden, The Netherlands, thus making available this collection of unique materials for the study of the history of Christianity in China. Mr. Lai's interest in bibliographical research and publication also led him to another major accomplishment at the Harvard-
Yenching Library: the publication in 1985-86 of the Library’s entire Chinese and Japanese catalogs in 72 volumes.

In the late 1930’s, Dr. A. Kaiming Chiu had contracted with the Harvard-Yenching Index Press in Peking to have the Library’s Chinese catalog printed in book form, and also to have individual cards printed for use by other East Asian libraries in the United States. The latter was 30 years before the concept of cooperative cataloging came into vogue among East Asian libraries during the 1970’s. The first three volumes of the catalog were published by 1940, and the rest set in type. Several thousands of individual cards were also printed and distributed to other East Asian libraries for use. Following Pearl Harbor, however, the Japanese troops went about destroying everything in Peking having to do with the United States, and the Harvard-Yenching Index Press located at the Yenching University campus literally went up in flames, and with it, the types that had already been set for the remaining volumes of the catalog. Dr. Chiu’s dream that the Library’s Chinese catalog would be published in its entirety was never realized in his life time. In the early 1980’s we finally made plans to have the Harvard-Yenching Library’s Chinese and Japanese card catalogs published through photo-duplication (The Library’s Korean catalog had already been published earlier). Put in charge of the project, Mr. Lai did all the preparatory work, including reviewing a very large portion of the catalog cards as well as adding a number of “see also” reference entries, and worked closely with G. K. Hall, the publisher, from start to finish. The 72-volume Chinese and Japanese book catalog, dedicated to the memory of Dr. Chiu, not only brought his dream to realization, it also stands as a testimony to Mr. Lai’s achievement.

Library automation came to East Asian libraries in the 1980’s. While moving gradually into an automated mode of operation, East Asian libraries had to grapple with the problem of how to integrate the manual catalog. We at the Harvard-Yenching Library explored the possibility of converting our card catalogs into machine-readable form, but found the cost prohibitive. Subsequently a $1.5 million grant from Harvard-Yenching Institute and a matching grant from Harvard University allowed us to contract out the conversion work to OCLC, a project which Mr. Lai oversaw. The retrospective conversion began in 1993 and was completed in 2000, four years after Mr. Lai retired. The total number of cards converted was some 360,000, the single largest database contributed to OCLC’s WorldCat at that time, undeniably a boon to the copy cataloging work of other East Asian libraries.

No tribute to Mr. Lai would be complete without mentioning his religious faith. Born into a Christian family, he has been a devout Christian all his life. It’s safe to assume that all the personal qualities he possesses are rooted in his Christian belief. Daisy Hu, a former colleague at the Harvard-Yenching Library, describes these qualities, quoting Confucius, as 溫良恭謙讓 (“gentle, kind, courteous, restrained and magnanimous”). This explains why Mr.
Lai is liked and respected so much by his library colleagues and friends. His Christian faith also led him to write profusely about the history of Christianity in Taiwan, particularly the Presbyterian Church in which he has served as an elder both in Taiwan and the United States. He also wrote a great deal of the history of Taiwan and is regarded as a leader in that field. There is a wealth of information about him on the internet. Two of the most informative sites are 賴永祥教授的學術生涯 (http://www.laijohn.com/life/Ng.IC%27view.htm) and 賴永祥长老史料庫 (http://www.laijohn.com).

As Mr. Lai is considering whether to return to live in Taiwan, I would like to pay this tribute to him that he richly deserves. I would also like to add a personal word: “Thank you, Mr. Lai, for your years of dedicated service at the Harvard-Yenching Library. Because of you, Harvard-Yenching became an even better library.”

August 2017
An Overview of Predatory Journal Publishing in Asia

Jingfeng Xia, Yue Li, Ping Situ

Abstract

Open access journal publishing has experienced a dramatic growth in the past decades and significantly reshaped the landscape of scholarly communication. In the evolving landscape, unprofessional practices have been discovered such as the so-called predatory publishing. Predatory journals are those that charge article processing fees to authors but fail to provide necessary editorial and publishing services (Berger and Cirasella, 2015). Most predatory journals are published in developing countries, particularly India and Pakistan (Ameen, 2017; Pulla, 2016). Our study attempts to provide an overview of this type of publishing in Asia, hoping to paint a fuller picture of open access by taking into account alternative practices. It finds that the predatory journals have primarily tried to benefit from the open access model to scam people by misrepresenting editorial information, promising a broad coverage of scholarly subjects, and levying charges for publications while ignoring peer review. It is our hope to use the findings to support academics in battling with unprofessional conduct.

Introduction

Studies find that the majority of predatory publishers are located in Asia – including continental Asia and the Middle East – particularly in India and Pakistan (Ameen, 2017; Bohannon, 2013; Lakhotia, 2015; Pulla, 2016; Seethapathy, Santhosh Kumar & Hareesha, 2016; Shen & Björk, 2015). India is also identified as a country where single-journal predatory publishers prefer to grow their business (Shen & Björk, 2015). Examinations of the origin of authors who publish in predatory journals singled out Asian countries, primarily China, India, and Pakistan, although Nigeria also ranks high on the list (Ezinwa Nwagwu & Ojemeni, 2015; Lakhotia, 2015; Xia et al., 2015).

This geographic distribution makes one wonder why this business model is Asia-centric and what effect predatory publishing has had on scholarly communication within Asia. The present study is an attempt to answer the questions by presenting readers with an overview of this questionable business. Specifically, the authors focus on exploring the routine business of predatory journals in some Asian countries, hoping to draw a complete picture of open access journal publishing by taking into account alternative practices.

Defined as unprofessional and lacking quality control, predatory publishing has many attributes that make it different from professionally managed scholarly publishing (Butler, 2013; Clarke & Smith, 2015; Raghavan, Dahanukar & Molur, 2015). Among many other qualities, a predatory journal is known for its article processing charges (APC), dubious editorial boards, and absence of a formal peer review process. Having taken advantage of the Internet technology, the open access movement, and the needs of young, inexperienced, or incompetent researchers for scholarly publications, avaricious individuals and publishers created shoddy websites where authors might be enticed to publish because of lax review, if any, and promised fast printing. Motivated primarily by hopes for quick financial rewards, opportunistic entrepreneurs engage in unprofessional exercises which have
created a negative impact on open access publishing in particular and on scholarly communication in general. In this ethical scenario, Asia is especially visible.

**Background**

Open access journal publishing, which is characterized by online availability with free content to readers, is an answer to ever-increasing subscription rates for scholarly journals, a continuous decrease in library budgets, and researchers' frustration at a lengthy and cumbersome publication process (Branin & Case, 1998; Laakso et al., 2011). The purpose of open access publishing is to remove price barriers, including licensing fees, subscriptions, and pay-per-view fees, as well as barriers such as copyright and licensing restrictions (Suber, 2015). The unrestricted access and reuse bring convenience and approachability to the latest research findings for everyone, resulting in immediate popularity for such journals.

In order to make open access journals accessible on a no-charge basis, publishing entities such as academic institutions and learned societies must provide necessary financial support for journal publishing and production. To make open access journals financially sustainable, a business model was developed that charges authors an article processing fee (Solomon & Björk, 2012; Van Noorden, 2013). Journals entailing such a charge carefully manage their review process to ensure that peer review decisions will be made based on quality. Many such fee-based open access journals have successfully maintained high scholarly vigor while retaining financial independence, such as the internationally recognized British Medical Journals (BMJ) and Public Library of Science (PLOS) journals. As of 2013, a total of 80% of all open access journals registered with the Directory of Open Access Journals (DOAJ) operated on the fee-based model (Kozak & Hartly, 2013).

Having observed the success of fee-based publishing, some businessmen found it a profitable business model and started creating their own open access journals. When profitability becomes the primary goal of publishing, scholarly quality is nearly always inevitably compromised. In order to maximize profit, such new journals needed to attract as many submissions as possible by making themselves look like legitimate journals. They did not hesitate to enhance their images by adding misleading words such as "global" or "international" to journal titles, promising publication within a week or two, and using fake journal impact factors. In recent years, such journals have mushroomed and established a notorious presence in the world of learned publishing (Beall, 2012a; Seethapathy, Santhosh Kumar & Hareesha, 2016).

The prosperity of such journal publishing reflects an imbalanced supply-demand market that sees an inadequate number of publishing venues and an increasing number of researchers who need publications for survival (Xia, 2014a). There are sociocultural and political factors that may have also driven submissions to these journals by authors from some developing countries, e.g., the culture that values quantity of publications more than quality, or simply policies that compel junior researchers, who do not usually have necessary experience and resources for conducting high quality scientific studies and completing scholarly acceptable articles, to publish in international journals (Pulla, 2016).

Facing these realities, scholars have devised strategies to battle these noxious practices. These include submitting grammatically correct but nonsensical papers to shabby open access journals to detect the deception (e.g., Bohannon, 2013; Davis, 2009; Djuric, 2015), and opening blog threads to expose particular journals (e.g., Eisen, 2013). Most famous are two long lists of the so-called
Predatory publishers and journals, amounting to a virtual blacklist of as many as 1,000 publishers and almost the same number of standalone open access journals (see Beall, 2016a, 2016b). These two lists were suddenly taken down in January of 2017 for unknown reasons, which shocked the media and scholarly community (Chawla, 2017; Silver, 2017). Fortunately, the Internet archive site Web Archive (web.archive.org) has preserved the lists, including all of their update history since the first publication of both lists.

Several studies have been conducted to investigate predatory journal publishing. For example, Björk and his colleagues published a series of articles seeking to understand it within the context of a scholarly publishing ecosystem (Björk, 2012; Laakso & Björk, 2012; Shen & Björk, 2105). Other scholars have engaged in empirical research to explore reasons behind such practices in selected countries (e.g., Ezinwa Nwagwu & Ojemeni, 2015; Nwagwu, 2015; Omobowale et al., 2014). Xia also has analyzed the article processing charges of a group of such journals (Xia, 2015) and has reviewed the quality of publications in two mega-journals (Xia, 2014b).

Regarding the origin of authors in predatory journals – based upon authors’ institutional affiliations, studies of randomly selected journals find that Asian countries, particularly India, Pakistan and China, are especially prominent (Seethapathy, Santhosh Kumar & Hareesha, 2016; Shen & Björk, 2015; Xia et al., 2015). The findings are strongly supported by a study on Nigerian predatory biomedical open access journals from 2007 to 2012 (Ezinwa Nwagwu & Ojemeni, 2015). A similar pattern also holds when it comes to the location of predatory journals based upon the contact information of their editor(s) or publisher, or currency used in transactions (Bohannon, 2013; Xia, 2015). These studies, however, have not narrowed down their subjects to Asian countries, even though case studies are available on the practice in one or two such nations (e.g., Lakhotia, 2015; Lin & Zhan, 2014; Mukherjee, 2014; Xia, 2014b). The present study seeks to fill that gap.

Method

Our data base is based on a list of standalone journals originally published on Scholarly Open Access (scholarlyoa.com) that documented questionable, scholarly open access journals identified as predatory. The data is now available through Web Archive which was recorded on February 3, 2017 for a list of journals last updated on December 30, 2016 (Web Archive, 2017). We understand that this list has drawn criticism, such as the ambiguous criteria of journal selection, as described in numerous blogs. However, it is the only comprehensive blacklist of its type and has been used in most, if not all, studies of predatory publishing. Most importantly, we found the listed journals to be really questionable in scholarly quality of their articles (Xia, 2014) and decided to use them for the analysis of this present research.

A simple random sampling strategy was applied to select a group of 300 titles out of a total of nearly 1,000 journals available as of December 2016. By consulting recognized criteria for determining predatory publishing on the same site, we identified types of publishing practices of interest to researchers and corresponding to those examined in earlier studies, e.g., journal locations, editorial members, article processing fees, subject areas, etc. To collect data, we manually opened websites and reviewed each of the selected journals.

Unsurprisingly, data collection was not an easy job. A common denominator for predatory practice is concealment of key information that might expose unprofessional and unethical conduct. As a result, for example, a journal’s location may be a P.O. Box in the United States although it charges...
Indian Rupees for article processing fees, simply because the journal seeks to create an international persona in order to attract submissions from a broader region. Whenever a journal's location was unidentifiable or not in Asia, it would be dropped off our sample list and would be replaced with another one picked at random. Thanks to the large number of Asian predatory journals, we had no trouble finding additional sources to keep our sample size at 300. With regard to article processing fees, we converted all currencies to U.S. dollars for the sake of comparison.

Due to time restrictions and the scope of our research design, we did not examine the scholarly quality of the publications. To achieve such a goal, alternative strategies would be in order. In another study by the senior author, the quality of predatory publications in terms of presentation (language use) and production (format) confirmed the shoddy quality of analyzed journals (Xia, 2014b). For the present study, we have availed ourselves of Scholarly Open Access that compiled a list of standalone predatory journals.

Findings

Geographic distribution

Regarding the geographic distribution of the predatory journals by country, our sample includes 16 countries or regions in Asia, among which India produces 235 journals, over 78% of all the total number. A distant second is Pakistan with 17 journals (about 6%), followed by Iran with 9 journals (about 3%) and Turkey with eight (about 2.7%). Other countries are limited to isolated cases (Figure 1).

Figure 1. Predatory journals in Asia by country

Compared to previous studies on the distribution pattern of predatory journals by country on a global scale (e.g., Xia, 2015), India is unrivaled. Yet its share among Asian countries is larger than its share among all countries in the world. This is simply because another predatory journal rich country, Nigeria, is not included in our analysis. On the other hand, although many Chinese publish in predatory journals, based upon their institutional affiliations (Ezinwa Nwagwu & Ojemeni, 2015;
Xia et al., 2015), China has only six such journals in our samples. We are witnessing the result of the Chinese government’s recently implementation of a set of harsh policies to regulate China’s publishing market, resulting in the suspension of many unprofessional journals (Lin & Zhan, 2014).

Editorial information
Although as many as 51 journals do not provide any information about their editors or editorial boards on their websites, 249 journals list an editorial board; some also have a review board. However, for some journals the same person serves both as chief editor and chief reviewer, with a shadowy editorial board. Not much information is provided for the credentials of the editor(s) or the editorial board members. In many cases, there is even no contact information at all for the editor(s). Gmail, Yahoo mail, Hotmail, or other free commercial email addresses are commonly used as the contact, compared to institutional affiliated emails associated with typical academic journals.

Publishing history and frequency
Most journals started in the early 2010s. Figure 2 reveals that 2012 was the peak year for the creation of journals. Only a year later, the number noticeably dropped. The market had apparently reached saturation. Even though a few journals launched as early as 2006 according to their websites, many provide no archival content until around 2012 or later.

Figure 2. Journal publishing history

The publishing frequency of each journal and the number of articles in each issue are very irregular. For example, Journal for Research, which started in 2015, has published two volumes so far, with 12 and 7 issues respectively. Volume 1 published seven articles in Issue 1, whereas the subsequent 5 issues (2, 3, 4, 5 and 6) published zero articles. Issues 8, 9, 10 and 11 each contained no more than five articles. International Journal for Innovative Research in Science and Technology is another example, where the number of articles varies very widely from 9 in one issue to 141 in another. The number of articles published presumably correlates with the number of submissions received, although we lack evidence to document that assumption.
Subject coverage
Predatory journals typically feature wide coverage of research subjects in order to maximize the number of submissions. This is precisely the case for journals in our study. Most of these claim to encompass all of the natural sciences, social sciences, or humanities, if not more. None is truly dedicated to a single well-defined subject area. Within our samples:
- 78 journals have multiple subject areas ranging from humanities and social sciences to agriculture/life science and computer sciences.
- 66 journals focus on phytomedicine, medical and biomedical sciences.
- 22 journals cover a broad range of research domains in engineering, as well as multifold aspects of sciences and technology.
- 14 journals accept submission in business, management and related subjects.
- 3 journals emphasize arts, along with many other broad subject areas that have nothing to do with arts such as agriculture.

Article processing charge
As shown in Figure 3, approximately 75% of the journals state on their website that they charge an article processing fee (APC). APCs range from $8 to $2,819. Most journals charge domestic authors their local currency while asking others to pay with US dollars, Euros, or another European currency - usually at a higher rate.

Some journals offer a fast track publication option if authors are willing to pay an additional APC. Many levy a mandatory excessive page/author fee in addition to their regular APC. For example, one journal charges 2,000 Indian rupees for a seven-page paper, plus 400 Indian rupees or $15 for each additional page. Another sets a standard rate of 1,000 Indian rupees for an article with up to two coauthors. For each additional author, an extra 300 Indian rupees are required.

A few journals claim not to charge an APC, but instead ask authors to make a monetary donation or purchase a subscription. For example, an annual membership rate of 500 Indian rupees is
required for Indian authors, or $100 for foreign authors. The *International Journal of Advance Research, Ideas and Innovations in Technology* asks authors to pay an online maintenance fee at the rate of 2,200 Indian rupees for Indian authors or $60 for non-resident ethnic Indians. For the few journals that provide no APC details on their websites, authors must communicate with the editor for the APC rate once their submission is accepted.

**Discussion**

Having outlined the general practices of predatory publishing in Asia, we may hazard some observations regarding (1) the geographic pattern of the practices, (2) misrepresentations of the publishing methods, and (3) purpose of the business.

It is clear that overwhelming majority of the predatory journals are published in only a few countries, especially in India and Pakistan, which two countries account for more than 80% of our samples. There are economic, cultural, and political reasons for this that deserve special investigation beyond the scope of the present study. Though accounting for the largest number of the unprofessional journals by country, India ranks only ninth in the DOAJ list with a total of 345 journals as of the fall of 2016. The DOAJ has implemented a strict screening process on the status of peer review for open access publishing when a journal registers with it. Therefore, although India boasts a highly profitable open access publishing industry (Mukherjee, 2014), there is a severe shortage of publications with quality and integrity. A previous study tries to explain the reasons by pointing to the great demand for young scholars in India to publish (Xia, 2014a). However, this is only one piece of the puzzle and falls far short when it comes to drawing a full-scale picture of digital scholarly communication in the country.

Pulla (2016) explains that India has developed an academic culture that evaluates the performance of researchers only by number of publications. Even top-flight research institutions have been pressured to chase the metrics. For example, a study finds a large percentage of publications in low-quality journals are authored by scholars from India’s premier government research bodies (Seethapathy, Santhosh Kumar & Hareesha, 2016). This academic culture has provided necessary soil for profit-oriented publishers to grow their business.

Predatory journals developed in India have not only allowed many domestic researchers to accumulate a satisfactory record of publications, but also successfully attracted authors from China for several reasons. In the past two decades, various internationalization policies developed and implemented by the Chinese government and institutions helped push researchers to submit their publications to open access journals based outside of China and published in English. Because English is an official language in India, their journals are typically in English, the standard language in scholarly publishing in the world. Also, publishing in an India journal is considered “international” by the Chinese standard. The quick and easy publishing promised by predatory journals seem to be an appropriate venue, and the benefit for academic advancement of Chinese authors outweighs the APC costs.

As briefly mentioned above, the situation in China is contradictory: there are few predatory journals published in China, but there are many China-based authors who publish in the predatory journals of other countries. In fact, China has suffered from a huge number of so called profit-seeking trash journals, as well as a flawed academic evaluation system (Lin & Zhan, 2014). However, this extraordinary practice has not attracted international attention because (1) they were mainly
published in Chinese and (2) the government was serious about “raising academic standards and weeding out slapdash and irrelevant publications” (Hvistendahl, 2011, p. 301).

Publishers purposely provide their journals with academic facades in order to attract submissions and maximize profits. Yet, they pay no attention to rigorous review and professional content. *Criteria for Determining Predatory Open-Access Publishers* provides a long list of checkpoints for identifying various wrongdoings in scholarly publishing (Beall, 2012b). Rather than repeat these points here, we will spotlight a new finding to support our indictment of unscrupulous behavior by predatory journals, that is, the presence of fake journal impact factors.

Journal impact factor (JIF) is derived from a formula that calculates the average number of citations for a journal on a yearly basis (Garfield, 1972). It is based upon journals indexed in the Journal Citation Reports (JCR), a product of Clarivate Analytics. Since JCR has a vigorous rule of selecting peer-reviewed scholarly journals and only a limited group of established journals are indexed in each academic field, open access journals, being a recent phenomenon, are typically not included. JIF has been internationally recognized as the standard in academia to verify the prestige of a journal and is, therefore, consulted when authors submit their articles.

It is not surprising that none of the predatory journals has been indexed in, or even considered by, JCR. This creates a market for fake JIF with predatory journals as target customers. Such businesses are mainly based in Asia and Africa. A typical fake JIF website delivers real JIF data, usually copied from JCR, to make it look legitimate and mixes with fabricated data for journals that pay for a JIF rating. The fake JIF provider may even make available an online certificate. By posting an eye-catching JIF value on its website, a predatory journal seeks to deceive authors and receive more submissions. Figure 4 is a copy of such JIF certificate linked on the website of the *Journal for Research*, one of our sample journals. Note that the JIF provider has a name, Scientific Journal Impact Factor, a shoddy imitation of the renowned journal impact factor.

In addition to being perceived, recorded, and assailed by academia, predatory publishing has attracted the attention of government authorities in different countries. In the United States, the Federal Trade Commission started to crack down on such dishonest publishers by filing a complaint against the OMICS Group in India. The commission is “seeking both monetary relief for researchers that have published with OMICS and to prevent the publisher from further violations of the Federal Trade Commission Act of 1914” (Straumsheim, 2016). China’s above-mentioned effort to block trash journals is another example. If the government of other concerned countries, particularly those prominent in India, could take similar actions, we would likely see a healthier scholarly publishing ecosystem.
Future studies may focus on (1) exploring the sociocultural, economic and political causes of predatory publishing in selected countries in the context of open access and scholarly communication, (2) examining both predatory publishers and standalone journals, and (3) developing strategies to reach an understanding of the behavior of the authors who publish in predatory journals. It is also important to distinguish predatory practices from unsophisticated forays into open access publishing.

Limitations
This present study collected its samples out of the predatory journal list developed by Beall. His criteria of predatory publishing has been criticized for relying too much on editorial looks such as defects of publishers’ websites, but ignoring systematic analyses of published content (Butler, 2013). As a review paper, we also did not present an examination of individual articles in studied journals on their scholarly quality. We simply relied on Beall’s judgement, which may end up with the inclusion of a few incorrectly classified open access journals.

Fortunately, scholars have explored the issue of predatory publishing for many years and analyzed predatory journals from various angles. It has become a recognized fact that the majority of journals on Beall’s list indeed pose certain levels of unprofessional practices in addition to their demand for article processing charges (Raghavan, Dahanukar & Molur, 2015; Xia, 2014b). Yet, readers are still advised to pay attention to the difference.

Conclusion
When we state that Asia is a major area in predatory publishing, we are, in fact, focusing upon a small number of countries, particularly China, India, and Pakistan, which have published an overwhelming portion of such journals. We have detected multifold varieties of professional misconduct by those “who are trying to benefit from the open access model to scam people” (Straumsheim, 2016). These
miscreants misrepresent editorial information, promise a broad coverage of scholarly subjects, and levy charges for publications. Armed with a better understanding of ways in which such practices damage and discredit scholarly communication, individuals in the academic community now should be able to work together, alongside government, institutional, and professional authorities, to battle this unprofessional business.

References


Current Situation and Challenges of Building a Japanese LGBTQ Ephemera Collection at Yale

Haruko Nakamura, Yoshie Yanagihara, and Tetsuyuki Shida

Historical background and current issues of LGBTQ groups in Japan

In recent years, Japanese activists for the rights of sexual minorities have begun to see a gleam of hope after decades of frustration with the slow pace of LGBTQ reform in Japan. In America, under the influence of the civil rights movement and the sexual revolution, the gay liberation movement centralized and organized after the Stonewall Riots in 1969, when members of the LGBTQ community fought back against routine police raids of their bars and community spaces. Japan also has a long history of neighborhoods, bars, and clubs for sexual minorities, but political organization and public visibility was limited. One early form of visibility was the publication of commercial magazines aimed at gay men like Barazoku, Sabu, and Adon which began in the 1970s in Japan. Also, Tōgo Ken 東郷健, who published a magazine called The Gay, also ran in national elections, but his activities remained officially invisible. Before this time, in terms of magazine media, Japan’s sexual minorities existed only in the borrowed corners of “kasutori” magazines, sex culture magazines focused on heterosexuals. Though gay men’s magazines date from the 1970s, lesbians had to wait until the 1990s for commercial magazines aimed at their community.

Japan had its first occurrence of AIDS in 1986, by which time the image of AIDS as a “mysterious illness often seen in gay men” had already been imported from America; gay men in Japan were now afflicted with the dual stigma of sexual deviance and illness. As discrimination against homosexuals and activism against homophobia increased, the Japan Association for the Lesbian & Gay Movement (NPO OCCUR) brought suit against the city of Tokyo in 1991, and this case, known as the “Tokyo Youth Home Incident,” became a symbol of the movement. OCCUR, a gay activist organization that had been established in February 1990 in Tokyo, had used a municipal youth home where members were subjected to discriminatory language by other guests and the director of the facility; ultimately they were refused permission to use the space by the director of the facility. They sued in 1991 and, in 1997, the plaintiff OCCUR won. In 1994, Minami Teishiro 南定四郎 became the organization head and the Tokyo Lesbian and Gay Parade was founded. After that, the Tokyo Parade changed lead organizers numerous times, but continues intermittently to today. Several metropolitan areas besides Tokyo, including Osaka and Sapporo, have also hosted successful pride parades.

Rights for transgender people have also expanded. In 1998, Sex Reassignment Surgery (SRS) was openly performed at Saitama Medical University Medical Center. Historically, SRS
in Japan had been restricted because of a case in the 1960s, known as the “Blue Boy” case, where there was a question of whether or not SRS was a violation of the Eugenic Protection Act; after that, the surgery could not be performed openly. In 2004, the “Act on Special Cases in Handling Gender Status for Persons with Gender Identity Disorder” was enacted, allowing people who have been diagnosed with Gender Identity Disorder to change the sex listed on their family registry and, after changing it, marry the “opposite sex.” This Act on Special Cases was accepted as good news by those experiencing gender dysphoria, though the law required those using it to meet particularly strict provisions, such as the complete destruction of reproductive capability, before their registered gender could be legally changed.

Despite increasing activism, Japan still does not recognize same-sex marriage, even though same sex marriage spread across Europe with the dawn of the 21st century, starting in Holland, and is now recognized in many countries. Certain local governments in Japan, however, have begun to recognize same-sex partnerships at the level of the district/city/town/village. This recognition began in Tokyo’s Shibuya and Setagaya areas in 2015 and after that, similar provisions were enacted in many local governments. This recognition by local governments does not guarantee legal rights between the two parties, nor does it constitute a binding arbitration legal contract, and it is difficult to find any significant legal benefits to recognition.

While systematic recognition of partnerships occurs sporadically, members of the LGBTQ community also turned to the issues of employment and child-bearing. For example, some nonprofit organizations have begun to hold seminars on LGBTQ employment aimed at graduating students. There are also members of the community who conduct lectures aimed at companies on how to create a positive work environment for LGBTQ employees. Higashi Koyuki 東 小雪, a former stage actress who became a lesbian writer and activist, was hired as spokesperson by an American surrogacy business aimed at gay men in 2016; she later welcomed a child with her partner Masuhara Hiroko 増原裕子. These movements can be seen as a sign that the LGBTQ community has made progress towards recognition. At the same time, there is strong criticism that many of these gains in rights have focused on assimilation into a lifestyle determined by the norms of the heteronormative society which marginalized the LGBTQ community in the first place. This criticism especially comes from scholars and those allied with queer theory.

New economic and political attention has also come towards the LGBTQ community from external sources. In 2012, Japan's major business magazines Weekly Toyo Keizai and Weekly Diamond reported that the domestic market for LGBTQ was 5.7 trillion yen and released a large special issue. Many members of the community viewed this economic interest with suspicion, especially with the focus on the LGBTQ community as a “money tree” and the association with Shibuya, which boasts the highest residential land prices in Japan (already globally considered high). Support has also come from political and global concerns: while preparing for the 2020 Tokyo Olympics, the Japan Olympic Committee (JOC) was pressured
to forbid discrimination based on sexual orientation and, in March 2015, the “LGBT Legislative Assembly” was formed, suggesting that concern with LGBTQ rights is spreading and expanding at the national level.

Rise of Japanese LGBTQ studies

While the LGBTQ community itself became engaged in political activism and the struggles of LGBTQ people were increasingly recognized by society, research on LGBTQ in the Japanese academy seemed to be stalled. According to Harada (2005), only an extremely small number of academic works with the keywords gay or lesbian had been published as of 2004.7 Shida Tetsuyuki also researched homosexuality (lesbian/gay, not including bisexual or transgender) in the field of sociology using search results from the Japanese National Institute of Informatics’s article database “CiNii.”8 According to this study, only a small number of researchers have published on homosexuality and those researchers tend to work on a variety of research topics, rather than specializing in homosexuality, preventing in-depth engagement with the topic. Thus, while there is much interest, the amount of research on LGBTQ topics is still limited. In recent years, the number of relevant dissertations by younger scholars has been increasing. This suggests a potential growth of researchers involved in research in this area. In 2002, Chūo University’s Social Sciences Research Center formed a research team on “Historical and Contemporary Issues in Sexuality” on the topic of phenomena related to LGBTQ. In 2008, the Japan Queer Studies Conference was formed, indicating increased interest in queer theory and promoting LGBTQ research.

Despite these signs of interest from academics, no Japanese university as of writing has a course or specialty field in LGBTQ studies. Undergraduate and graduate students, for the most part, must learn under the guidance of researchers on gender who are also knowledgeable about sexual minorities. In 2015, the Japan Queer Studies Conference mentioned above ceased activity indefinitely after a conflict between academic members and activist members. On the other hand, Waseda University has recently expanded its interest in LGBTQ research and support, hiring for a Queer Studies position and creating an LGBT student center, the GS Center, based on a student proposal.9 While Japanese universities lack much of the infrastructure which would support research and education on sexual minorities, it is clear that interest remains and is growing amongst the younger generations of scholars, community members, and activists.

Collecting Ephemera: Academic value as scholars

Ephemera is at particular risk of loss and disposal without the intervention of archival preservation and institutional collection. Material which was assigned an ISBN is collected by the National Diet Library as a matter of course, so commercial magazines for gay men such as Barazoku, Sabu, and Adon are accessible through the NDL and at archives such as Ōya
Shōichi Bunko. However, similar resources for ephemeral materials do not exist, making these materials difficult to locate and preserve. Collections of LGBTQ ephemera held by private individuals are often disposed of by family members after the individual’s death, particularly in cases where the deceased was closeted. We estimate the amount of ephemera at risk of disposal in similar situations is significant. Some Japanese academics are also aware of this difficulty. By directing institutional resources towards the archiving of ephemera, we can identify private collectors and preserve their collections.

Although not widely available, ephemeral materials give scholars a glimpse into the time and space of a community in unique ways that cannot be replicated by other, more traditional archive materials. This is of particular importance for groups like the LGBTQ community, which exist as separate subcultures and which are often overlooked both by popular mass media and large archives centered on heterosexual society in Japan. For instance, Ishida Hitoshi’s research analyzes so-called Inran Ryokan 淫乱旅館, hotels or bathhouses which served as cruising spots for gay men. The paper refers to advertisements in magazines which show that ephemeral materials such as fliers would be helpful for this sort of research. Another example is a study which analyzes magazine articles and contributions from readers regarding people’s sentiments towards faux marriages. It mentions an advertisement for sham marriage arrangements made between gays and lesbians posted by the editor. Although this particular study uses traditional magazines to analyze the topic, this sort of study can utilize more ephemeral materials and gray area literature since the topic could be considered controversial and unethical for the heterosexual community. Another case deals with activities related to women’s liberation in the late 1970s. The study analyzes three community newsletters published in the 1970s, materials which, as independent community publications, fall under the umbrella of ephemera. The author analyses those magazines as tools for the first movement to try to dispel negative connotations of lesbianism. There is particular academic need to collect ephemeral materials regarding lesbian communities, because the lesbian book market is smaller than the market for gay men and because lesbian-related publications are more often informally published materials, such as non-commercial newsletters.

Western scholarly works on Japanese LGBTQ have used a variety of sources. For instance, Queer Voices from Japan by Mark McLelland introduces voices of sexual minority groups in Japan by translating personal memoirs and interviews, making them available to English-speaking scholars. This work shows the value of archiving personal interviews, diaries, and letters as significant primary sources. In Lesbians in East Asia, sources like ECQA (Ecumenical Community for Queer Activism in Japan) newsletters and policy documents were used for research on lesbian rights movements in Japan. Another chapter in the book discusses the process of creating an oral history of lesbians in Hong Kong. McLelland’s Queer Japan from the Pacific War to the Internet Age also cites minor magazines published in Japan, promotional fliers of night clubs and films, and internet home pages of ordinary people.
LGBTQ collections in U.S. and Japan

Initial efforts to collect LGBTQ-related materials in the United States were connected to the social justice movements of 1960s and 1970s for people of color, women, and gays and lesbians. The Stonewall Inn riots in 1969 are dated as beginning of modern gay political rights movement. In 1970s, community members formed archives for the preservation of LGBTQ history, such as the Lesbian Herstory Archives and the ONE Archives. By the 1980s, the AIDS crisis precipitated greater political organization; the focus on collecting materials and studying LGBT communities acquired an increasing sense of crisis—the feeling that this material will be lost if the community itself does not begin to archive it. Finally, increasing mainstream acceptance brought greater political and economic power from the 1990s to the 2000s. As these movements progressed, LGBTQ issues and identities also diversified legally, culturally, and socially. Interest in academic studies related to LGBTQ subjects grew, necessitating the preservation of the voices and history of LGBTQ people. With this historical and social background, LGBTQ materials are collected in repositories like the university libraries of Cornell and New York University; the number of archivists specializing in the topic grew, resulting in the formation of The Lesbian and Gay Archives Roundtable (LAGAR) by the members of the Society of American Archivists. Because of the strong interest in LGBTQ studies at Yale, the University Library has also focused on collecting, organizing, and preserving LGBTQ-related materials in various formats in the United States.

We conducted a brief survey of LGBTQ-related collections in Japan by contacting several research institutions that are known to be active in related research. The survey showed that no large LGBTQ collections existed at either major public or university libraries. Usually LGBTQ materials are collected only as a part of gender studies or women's studies collections by public institutions that specialize in women and gender issues such as the Center for Gender Studies at International Christian University, the National Women's Education Center of Japan, or the Osaka Prefectural Center for Youth and Gender Equality. However, these same initiatives also indicate that social awareness of LGBTQ is spreading in the field; the Center for Gender Studies at International Christian University, for example, an employee intentionally labels books related to the subject of sexual minorities and the Osaka Prefectural Center for Youth and Gender Equality conducted public relations activities for LGBTQ groups and related social issues in 2015.

If one looks outside the sphere of public research organizations, some institutions are working on gathering small LGBTQ collections including ephemeral materials. Like early US efforts, however, initiatives for collective archives are community driven. According to Ozawa (2014), there are many libraries operated by the LGBTQ community itself. A representative example is LOUD, an organization which designates itself as a “space for lesbian and bisexual women.” Other examples in recent years include collections related to sexual minorities at NPO SHIP Nijiiro Cabin in Kanagawa, Ehime LGBT Center Niji-kara Space in
Ehime, and Osora ni Niji o Kakemashita in Aomori. These libraries often include ephemera, but unfortunately libraries operated by members of the community are often beset with economic difficulties and operate under constant threat of closure. Under these conditions, Ozawa asserts the necessity of archiving and intentionally collecting publications and documents written from the perspective of community members, starting with newsletters published from within the community, in order for minorities to construct “a self-sustaining history,” to form a certain type of “imagined community,” and to create a common space of belonging for community members and allies. There are also some private collectors and community members with personal collections. However, these personal collectors are often skeptical of institutional-affiliated collectors, and their collections are very difficult to access from outside of the collectors’ communities.

In Western countries, a few organizations are collecting Japanese LGBTQ related ephemera. ONE National Gay and Lesbian Archives at University of Southern California currently focuses primarily on American LGBT history, but the archive director, Dr. Joseph Hawkins, is trying to build an international collection, including Japanese materials. Their collection includes the entire run of Barazoku and other gay interest magazines. The archive director’s personal collection of Japanese materials, accessible in his office by appointment, includes additional gay interest magazines, photocopies of 3000 documents on transgender-related materials held by Chūo University, and fuzoku zasshi from the 1920s to the 1950s. The University of Melbourne in Australia also holds a collection which consists of “hard-to-find Japanese materials dealing with sexual minority cultures and identities across the twentieth century.” The archive, led by Mark McLelland, produces a large amount of scholarship related to Japanese LGBTQ issues. The project also has a downloadable and searchable spreadsheet with detailed bibliographies of the collection, which serve as an introduction to significant resources on Japanese LGBTQ-related studies.

Why did we started collecting?

Yale started collecting Japanese LGBTQ-related materials in 2004 for several reasons. First, Yale University has played a leading role in the development of LGBTQ studies for more than thirty years, starting with hosting large conferences in 1980s. For example, Larry Kramer, celebrated playwright, writer and activist for gay rights in the United States, donated his personal collection to Yale University in 2001. The collection consists of “manuscripts and drafts of plays, books, screenplays, and articles; AIDS-related material; diaries; correspondence; photographs; printed material; audiovisual material; and other papers. There is also substantial material relating to the founding of Gay Men’s Health Crisis and the AIDS Coalition to Unleash Power (ACT UP) in particular, and to the AIDS movement in general.” Along with this donation, Yale received a donation of 1 million dollars for the creation of the
Larry Kramer Initiative for Lesbian and Gay Studies at the University. The initiative supported research and teaching in LGBT studies, as well as a five-year full-time executive coordinator position. 22

With these strong academic interests within the University, it was natural move for the Yale anthropology department to welcome a new faculty with Japanese LGBTQ-related academic interests in 2004. Professor Karen Nakamura, the author of many works related to gender and sexuality in Japan started to offer several gender and LGBTQ-related courses and attracting graduate students with similar interests. 23 This prompted the library to begin collecting related Japanese materials as well as archival collections.

As for resources related to Japanese LGBTQ, however, it has been challenging to either locate or access the majority of them in the United States. Although the lack of accessibility of Japanese LGBTQ materials motivated the library to consider building a substantial collection, starting to build a new collection, especially archives, requires the assessment of possible challenges including the issue of sustainability. Haruko Nakamura, the librarian for Japan Studies at East Asia Library at Yale University, had been collecting Japanese LGBTQ-related publications up to this point. She saw an opportunity to build an archival collection in order to document Japanese LGBTQ communities for scholarly research, as well as attempt to broaden LGBTQ studies in U.S. beyond North American and European communities. After considerable evaluation and planning, originally initiated by Haruko Nakamura, the library decided to develop a Japanese LGBTQ collection; several factors contributed to this decision.

The first major factor was the excellent support and infrastructure for the collection of related materials within the library. As mentioned above, with strong academic interest and funding from the university, the library already had a long history of LGBTQ collection development and experience with collecting LGBTQ materials, including archives. The existing library LGBTQ collection supported the creation of a Japanese collection with funding, and advice and training from expert archivists at the Manuscripts and Archives department where the majority of LGBTQ-related collections reside. Furthermore, the Manuscripts and Archives department is well-equipped to provide access to the collection with a secure reading room and reference services.

As the first step, the East Asia Library started collecting traditional library resources such as books and journals, including the entire run of gay interest magazine Barazoku. A research guide for this topic was also created to support discovery of those materials. 24 In addition, since Yale University had been actively collecting resources related to this subject, the library had already developed the local subject heading LGBTQ resource and applied it to all acquired LGBTQ materials. By entering the term “LGBTQ” as a local subject heading or keyword and limiting the search to Japanese, library users are able to find all books and journals related to LGBTQ issues published in Japan. In March of 2017, Yukari Sugiyama, the Japanese cataloger at Yale University Library, ran a WorldCat Search API to analyze LGBTQ related collections held in libraries in North America. 25 Through the search and its findings, we were able to see the uniqueness of the LGBTQ collection at Yale, as well as which missing titles we
could acquire. As of August 2017, there are more than three hundred related Japanese LGBTQ resources at Yale, including books, audio visuals and journals such as *Bara, Ronso Kwia* = *Journal of Queer Studies Japan* and *Mlmw*. However, those types of subject headings can be problematic, with ambiguous definitions of LGBTQ and fuzzy subject parameters which overlap not only with gender and sexuality issues, but also with larger social, political, and cultural topics. Thus, while books are labeled as part of the LGBTQ collection at the librarians’ discretion, researchers are advised to be cautious and understand that the library may hold additional related materials.

**Collecting Ephemera: challenges of communities and collecting**

From the first stage of collecting LGBTQ-related ephemera and archives from Japan, the library immediately faced several challenges. Unlike commercially available publications, those materials are not distributed widely and figuring out sources of material and their availability was the first challenge. The initial approach to LGBTQ communities by the library caused some suspicion due to the often-closed nature sexual minority communities.

Yoshie Yanagihara was visiting Yale for one year through an exchange program between Yale and the University of Tokyo. When Nakamura shared the challenges of collecting Japanese LGBTQ materials with Yanagihara, the discussion led to a formal project to establish procedures to collect archival materials operated in two locations, in Japan and at Yale. Because of his subject knowledge and long experience working with LGBTQ communities in Japan, Yanagihara also invited her colleague, Dr. Tetsuyuki Shida, to this project. In order to secure funds for both Yanagihara and Nakamura to travel and collect materials in Tokyo, work contracts as non-Yale employment, payment procedures, and shipping methods were determined. The team to work on the Japanese LGBTQ ephemera at Yale University therefore consists of Haruko Nakamura, Yoshie Yanagihara and Tetsuyuki Shida.

As for collecting strategies, the team determined that there were roughly two types of LGBTQ-related archival materials. One is freely available publications from major LGBTQ support organizations. The other type is independently produced items like fliers and leaflets, which are harder to collect at regular intervals. For this sort of material, we first listed those organizations producing related ephemera. In order to obtain these materials, we first approached an outreach organization, Community Center Akta, to help us by sharing the fliers they distribute in their community space. Akta kindly agreed to send those materials to us regularly for a small fee as a donation, which reduced the time and effort of collecting these fliers from individual places. Akta is a non-profit organization established to bring awareness to HIV/AIDS. It is a community center, located in the biggest gay town in Asia, Shinjuku 2-chome and also functions as an information hub for LGBTQ people in Japan. They display both their own and other organizations’ flyers and leaflets, covering many aspects of LGBTQ life as well as HIV awareness.

Considering various expenses, including transportation costs, the small fee to Akta, as
well as the “consultation” fee to Yanagihara and Shida, the library budgeted $200 dollars per month for the project. The budget, however, does not include costs for archival supplies, student workers’ wages for processing materials, as well as other hidden costs such as space for storage. However, we are able to cover the cost as part of the general operations of the East Asia Library or as a part of the expense for the Japanese film ephemera collection.

As mentioned previously, even with the help of Dr. Yanagihara and Dr. Shida, the major challenge to accessing LGBTQ materials has been closed and skeptical attitudes of community members toward those outside the LGBTQ community. For example, one of the team’s acquaintances in Japan has a massive collection of over 10,000 items. When we asked him if he would be willing to sell or donate the collection, he first expressed his willingness to consider these options as a personal favor to a friend. However, after he found out that our purpose was to build a public archive outside of Japan and most of us working on this collection were not members of the LGBTQ community, he declined to contribute. He expressed his willingness to contribute his collection to younger generations of sexual minorities in Japan. He also felt some discomfort with sharing his archive with those outside of the gay community.

From this sort of experience, we learned that there are multiple reasons why collectors do not support the project. One of those reasons is the struggle of being part of a minority group in heterosexual world. The stigma of being gay can make them very cautious about people outside of the gay communities. The other issue is mistrust of US institutions. Having the project based at Yale and the US institutions seemed to trigger negative ideas of American centrism. Japanese collective trauma may cause impressions that “Yale” or more widely “the US” haughtily robs them of their materials and misconstrues them in terms of Western cultural value systems, ignoring the actual Japanese contexts.

We also found out that it is more difficult to collect materials outside of the Tokyo metropolitan area. As with Akta, we planned to cooperate with another organization in Osaka. However, when speaking with the director, he expressed his fear of giving the materials to the public due to privacy concerns. Regardless, this was not detrimental to the project as most of these materials would have overlapped with the materials collected in Tokyo.

Another issue raised was the consideration of accessibility and privacy. A few donors have expressed some concern about privacy. Even though fliers and brochures were created for public distribution, they were initially intended for the only local and select communities, creating a reasonable expectation of privacy. Once these ephemera are placed in the repository, they can be accessed and studied outsider of the LGBTQ community in foreign countries. Some see this as unnecessary exposure. The Yale Japanese LGBTQ collection is still small and relatively unknown. Still, the fear of being exposed of the members of the LGBTQ community has been greater than the collecting team expected. This sort of issue will grow if and when personal manuscripts are collected. In the future, therefore, staff who process new acquisitions for this collection may have to sort certain items as restricted to the public.
until a later agreed-upon disclosure date.

Archival communities in the United States have been dealing with privacy issues in several ways. If a collection is acquired as a personal donation, a gift agreement with a document restricting access to the archive is negotiated between the institution and the donors to outline the length or end date of restriction. The David Benjamin Mixner papers held at Yale is an example of this case. David Benjamin Mixner is a well-known activist, consultant, and author in the United States. This archive collection documents his life as a leader in the gay rights movement, author, and political consultant and advisor. The finding aid to this collection provides the following statement about restriction to access: “The papers are closed until January 1, 2031, unless researchers receive permission in writing to access them from the donor, the person holding power of attorney for the donor, or the executor of the estate of the donor. Researchers wishing to request access should email mssa.assist@yale.edu requesting specific box numbers in order to initiate the permission process, which may take several weeks.”

Ephemera collections in general are slightly different from archives like organizational records or manuscripts, since they are publicly created and distributed. Reproduction of materials in a certain collection can be restricted by inserting specific wording in a finding aid. For example, the records at Yale of the John Hay Whitney Foundation, a nonprofit organization dedicated to education and social welfare, have a restriction on reproduction in the finding aid, stated as, “Reproductions in any form of this material require the written permission of the Greentree Foundation until January 1, 2050. After January 1, 2050, reproduction is permitted but subject to applicable copyright laws.” Secure reading rooms also provide a layer of security for physical materials: users are required to register as a reader and sign the terms of use including abiding copyright law.

Collecting Ephemera: challenges of processing and organization of materials

Despite those challenges, the Japanese LGBTQ-related ephemera collection has been slowly yet steadily growing. Types of materials collected are advertisements, community newspapers/newsletters, and pamphlets. Currently, the ephemera collection consists of four archival boxes of those ephemera materials. While the materials were being collected, however, the library also faced the difficult challenge of organizing and processing ephemeral materials. The first challenge was to determine the organization of the collection. An arrangement scheme had to be comprehensive and orderly, yet inclusive enough to allow the collection to be expanded in the future. It should also be clear enough for users to find materials easily. Initially, categories by creation date, formats, LGBTQ groups, or subjects were considered as the possible organizing schemes. Each of these, however, contained particular challenges and issues for clear categorization and it was impossible to avoid overlapping
categories. Because of their extensive experience and expert knowledge, the Japanese librarian sought the guidance of archivists from the Manuscripts and Archives Collection in the library.

To avoid as much confusion as possible, extra content notes were included to clarify the scope of potentially overlapping and confusing categories. The arrangement and descriptions also attempted to anticipate how users might use the collection based on how particular academic disciplines are formed as well as what type of material forms the majority of the collection so far. In the end, we decided to categorize the collection by both subjects and format. Because the collection included significant newsletters by LGBTQ-affiliated organizations, these newsletters were arranged in their own series. The rest of the material consists of fliers and brochures sorted by subjects with subcategories. (Figure 1)

Figure 1: Yale Japanese LGBTQ ephemera organizational scheme

<table>
<thead>
<tr>
<th>Series 1: Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Academia</strong></td>
</tr>
<tr>
<td>Lectures, studies, classes related to academic study of LGBTQ experience.</td>
</tr>
<tr>
<td><strong>Advertising</strong></td>
</tr>
<tr>
<td>Advertising for products targeted at LGBTQ community but not necessarily from LGBTQ businesses, including dating sites, escort services.</td>
</tr>
<tr>
<td><strong>Arts and Culture</strong></td>
</tr>
<tr>
<td>- General – Includes theater, music, painting, photography, and literature</td>
</tr>
<tr>
<td>- Film</td>
</tr>
<tr>
<td><strong>Community</strong></td>
</tr>
<tr>
<td>Spaces and activities for the purposes of LGBTQ community building.</td>
</tr>
<tr>
<td>- General - Community centers and social events (except those at bars and clubs), friendship events, religious meetings</td>
</tr>
<tr>
<td>- Bars and Clubs</td>
</tr>
<tr>
<td>- Pride events</td>
</tr>
<tr>
<td><strong>Family</strong></td>
</tr>
<tr>
<td>Includes Marriage, Adoption, Reproductive Technologies.</td>
</tr>
</tbody>
</table>
Once we determined the scheme for sorting the collection, we started physically arranging and describing the materials in the finding aid for the collection. First, the library acquired acid free folders and boxes approved by the Manuscripts and Archives Collection. A student assistant was hired and trained to physically process the collected materials according to the decided scheme. We had additional discussions to modify the sorting scheme as the student sorted through materials. This process was expected, as the East Asia Library already had experience of processing the Japanese Film Ephemera Collection. Because we physically sort materials at the final stage, we usually face some sort of discrepancies in sorting schemes which require us to constantly adjust and modify the scheme. Although those processes may seem redundant and somewhat backward, this careful process to develop with a coherent and logical scheme will enable us to add to the additional materials and to expand the archives almost indefinitely.

For the actual creation of the finding aid, we first created an Excel file with an itemized list of files in the collection with various information such as a summary of the collection, content description, provenance, copyright and citations guidance, etc. The Excel data was converted to EAD and imported in ArchivesSpace. ArchivesSpace is an archival managing system for creating and publishing finding aids. It is an open source platform which was adapted and customized to fit Yale University Library needs in 2015. Since ArchivesSpaces is difficult to operate for student workers without extensive training, the East Asia Library was only responsible for creating the Excel file with the itemized list. Once the detailed list was
completed, the converting of the Excel file was done by staff at Manuscripts and Archives.

After all materials were sorted and stored in files and boxes, the collection was transferred to Manuscripts and Archives, who finalized and published the finding aid, and created a catalog record. Access to the collection is managed by the Manuscripts and Archive department, who has the funding and staff to provide secure access and ancillary services. The presence of staff in the reading room can also provide extra security and any assistance such as reference questions and photo copy services, as well as easily handled remote requests such as digital/photo copy. The library completed for the first batch of collection in December 2016 and was made available to the public starting January 2017.2930

Figure 2: The first page of Guide to the Japanese Lesbian, Gay, Bisexual, Transgender, and Queer Collection (MS 2067). Manuscripts and Archives, Yale University Library.
Expansion and future projects

As mentioned above, the first batch of the collection was processed with a detailed finding aid and deposited in the university’s Manuscripts and Archives which provides access to the collection in a secure reading room with services such as copying and digitization. While the size of the collection will be expanding as more materials are added, another interesting area of collection development for this sort of collection will be archiving and data mining of born digital content. For instance, Regumi is the one of largest lesbian organizations in Japan and its website contains information related to their activities and support for lesbians. Over the years, however, the website has had constant additions and deletions as well as undergoing some restructuring of the site itself. Fortunately, their old website was archived by the Internet Archive and is available for researchers to view via the Wayback Machine. Because of these archival efforts, researchers can see the transitions in their activities and growth through their website.31

Several departments of Yale University Library have already started to collect their own university websites including social media like blogs and tweets and some curators have decided to collect websites which complement subjects the library has historically collected. For example, Yale University Library is a member of a project called Contemporary Composers Web Archives. This project was formed by several university libraries collectively known as the Borrow Direct Music Librarians Group. The project attempts to “to preserve copies of present and future manifestations of the websites of notable contemporary composers in a secure digital archive to guarantee the continuing availability of these extremely important but potentially ephemeral documents for researchers and scholars seeking to study the careers of contemporary composers.”32 This project is an excellent example of a collaborative project among several university libraries to share the possibly enormous cost of web archiving, including subscriptions to web archiving tools like ArchiveIt, staff time for the constant monitoring of archived sites and other tasks, and acquiring permission to archive from web content creators. Furthermore, this sort of collaborative project also encourages participating libraries to discuss and promote standardization of various policies regarding collection development, use of copyrighted contents, and permissions for digital archiving projects.

Once the processes for collecting their archives are standardized, those projects can be extended further by promoting use of the archived collections. With digitally collected archives, researchers can also approach new ways of intellectual interpretation through data manipulation and creating digital scholarship. For example, The Yale Community Voices Archive (YCVA) project aims to collect social media and born-digital content on Yale campus discussions of diversity, allowing a new perspective on current campus climate via digital archiving. Similar projects could be used to track changes in conversations around LGBTQ issues and identities in Japan.
Conclusion

This paper attempts to introduce the academic importance of collecting Japanese LGBTQ-related ephemera as well as the process of collecting, organizing and providing access to the collection at Yale University. By doing so, it introduces several distinct characteristics and challenges of collecting ephemeral materials produced by or associated with Japanese sexual minority groups. Those characteristics can be traced to the rarity of archives and sources of material and the privacy concerns of sexual minority groups. However, numerous other religious and racial minority groups create ephemeral materials which may share similar characteristics and research importance. While this paper only focuses on collecting ephemera associated with Japanese sexual minority groups, some of those practices may also be utilized to collect other archive collections associated with minority groups and subcultures. Hopefully this paper raises awareness of the importance of such collections while also providing some practical methods and creative ideas for collecting similar archival collections.
As widely known in the United States, the term LGBTQ stands for Lesbian, Gay, Bisexual, Transgender and Queer. Q can also refer to Question, implying that the LGBTQ-related collections discussed in this paper include any materials related to sexual minorities.

Haruko Nakamura 中村宏子 is the Librarian for Japanese Studies at Yale University Library. Yoshie Yanagihara 柳原良江 is currently an assistant professor at Tokyo Denki University as a sociologist. Her area of study is Gender Studies and Cultural Sociology. Tetsuyuki Shida 志田哲之 is a part time lecturer at Waseda University as a sociologist. He specializes in Sexuality Studies and Family Studies.

The paper treated the Stonewall Riots as the flashpoint in the modern gay liberation movement in the United States. While outlining the history of the LGBTQ liberation and civil right movements is out of scope for this paper, we consulted some sources relating to the earlier gay civil right movement in the United States such as D’Emilio, J. Sexual Politics, Sexual Communities, Chicago: The University of Chicago Press, 1983; Faderman, Lillian. The Gay Revolution: The Story of the Struggle. New York Simon & Schuster, 2015; and Caldera, Mary, "The Lesbian in the Archives: An Overview of the History," in Perspectives on Women's Archives (edited by Tanya Zanish-Belcher with Anke Voss). Society of American Archivists, 2013: 215-246. This paper treated the Stonewall Riots as the modern and flashpoint gay liberation in the United States.


Those institutions include Chūō University, Wakō University, Jōsai University, International Christian University, Tokyo Women's Plaza, Kokuritsu Joseikyoiku Hall and Dawn Center - Osaka Prefectural Center for Youth and Gender Equality, etc.


Many thanks to Dr. Hawkins for providing the information about Japanese collections at ONE. Before becoming director/archivist at ONE, Dr. Hawkins did anthropological research in Japan, thus he has extensive knowledge of Japan and the problems of collecting Japanese materials.


Larry Kramer Papers. Yale Collection of American Literature, Beinecke Rare Book and Manuscript Library.


This project was presented as “Data-Driven Collection Analysis Using the WorldCat Search API:Japanese LGBTQ Collections” in Advancing Digital Scholarship in Japanese Studies: Innovations and Challenges, University of Toronto, March 13-14, 2017

The collection is mostly composed of printed materials on Japanese films produced by Japanese sources. Its main contents include handbills (chirashi), press kits, theater pamphlets, film newsletters as well as catalogs, all dating from approximately 1970 to the present. The collection, which is still expanding, currently covers over 600 contemporary Japanese directors and their films along with major film festivals and film events in Japan. It offers a comprehensive resource for researchers investigating the breadth of contemporary Japanese film and visual culture or those studying individual films and filmmakers. The collected materials broadly document the great diversity in contemporary Japanese film and its culture and represent types of materials often overlooked by academic collections and archives. Therefore, researchers with a broad range of interests, especially Film and Literary Studies, Cultural Studies, Journalism, Sociology, and Art History, may be interested in the collection.

David Benjamin Mixner Papers (MS 1862). Manuscripts and Archives, Yale University Library.


Special thanks to Mary Caldera, Head of Arrangement & Description at Yale’s Manuscripts and Archives Collection for her guidance in building this collection and producing the finding aid as part of their database and to research assistant Caitlin Casiello for her hard work on creating the finding aid for this collection, as well as her insight and advice on writing this paper.


Using Data Visualization to Examine Translated Korean Literature

Kyung Eun (Alex) Hur and Hyokyoung Yi

Introduction

In May 2016, the English translation of a South Korean fiction book called The Vegetarian won the Man Booker International Prize, one of the world’s most prestigious literary awards. The Vegetarian was Han Kang’s first novel to be translated into English by Deborah Smith, a British translator. The work was also chosen as one of the NY Times Book Review’s “The 10 Best Books of 2016.” It is considered translated Korean literature’s biggest win since Kyung-Sook Shin’s Please Look After Mom won the closing Man Asian Literary Prize in 2012.

Korean literature is becoming one of the major Korean cultural icons for global citizens, joining the Korean waves of drama and K-POP. The translation of literary works into other languages is a key factor in connecting people with different languages. In light of the increasing attention to the translation of Korean literary works, the idea emerged to gather data on the Korean literary works that have been translated into other languages to better examine the historical and current trends of Korean literary translations.

Many print publications of the compilation of translated works exist (Han’guk Munhak Pŏnyŏgwŏn, 2002; Kim, 1998; Kim, 2012; Park, 1993), and the Literature Translation Institute of Korea (LTI Korea) Library offers an online bibliography of the Korean literature translated into various languages on its website (http://library.klti.or.kr/bibliography). However, these thick reference volumes do not provide a quick and easy snapshot for understanding the whole picture of the translation trends in Korean literature. Although the online bibliography does provide some easier access to translated titles by language, publication year, country, and genre, neither resource provides visual access that can be grasped quickly. They also do not have the interactive functions that users can use to view, sort, and rank the information in a visually presentable way.

As data analysis and visualization have been rapidly introduced in many research areas, including humanities, this project’s goal was set to visualize the translation data of Korean literature using Tableau, an interactive visualization tool, to better examine the topic. The study also aimed to discover any new findings or hidden information through data analysis with the help of new technology. Finally, the study hopes to bring a new perspective to the topic so that users can dig in for more information on their own.
**Methodology**

**Data Source**

In order to create an accurate and meaningful visualization, it was very important for us to carefully select accurate and reliable data. We began the data-selection process by looking through credible online open sources that hold catalogued data of translated Korean literature, such as OCLC WorldCat, National Library of Korea (국립중앙도서관), and LTI Korea (한국문학번역원). Manually collecting all bibliographic data from these open sources was not possible. For example, OCLC WorldCat offered no freely available option to search for all the books translated from Korean. Thus, we contacted the other open source organizations to get data on the translated books. However, obtaining such data from OCLC WorldCat and National Library of Korea was not successful. OCLC WorldCat’s data service was very costly, and National Library of Korea’s data were lacking in terms of the amount provided. In the end, we decided to go with the data provided by LTI Korea, which did not involve costs but had a comprehensive number of bibliographic records for the use of our study.

**Data Description**

LTI Korea’s catalogue had a total of 4,526 book records. Each book's record included metadata on the author, publication date, original title, translated title, translated language, genre, and translator. In order to improve the dataset, we also manually added data on the original year of translation to determine the translation time for the books. Our target literary period was from the Korean pre-modern classics to modern literature. We filtered out non-literary materials, such as biographies, self-improvement books, children's literature, and books with multiple or unknown authors. In terms of genres, we included in the dataset fiction, poetry, soap opera scripts, and diaries/journals. The exclusion was made to make the data more consistent and relevant to each other for comparison and analysis. After filtering out the unnecessary data, we were left with a total of 2,172 books. These data were used for this project.

**Data Processing**

We pre-processed the finalized data to ensure data quality by making sure that the information was complete, consistent, and accurate (Lemieux et al., 2014). Much of the data, especially the titles and years of translation, was missing, inaccurate, or inconsistent. Thus, it took some time for thorough data cleaning and editing. In this step, we made sure that the data did not have any missing values or inconsistencies. We also conducted discrepancy
detection to ensure that there were no outliers or errors in the domain of each attribute. Finally, we excluded the inconsistent use of numbers, names, and data representations.

**Data Visualization**

After pre-processing the data, we sought the best possible way to visualize the data. We decided to use Tableau as our tool for various reasons. First, Tableau offers a tabular nature for viewing data. Second, it is the most recent and up-to-date interactive visualization tool that is flexible and fast. Third, Tableau was freely available to test and had all the necessary visual presentation features that we wanted to use for our study. In addition to the public nature of this program, which made it easier to access by anyone who wished to, we were able to get the Tableau desktop version for free so we could freely test the data for visual presentation. We also wanted to create a way for users to easily view and understand the data. Tableau fulfills this goal with its fast interactivity and good user interface. Tableau enables the quick transformation of data into a picture, allowing for various types of visualizations. It also has the ability to make the visualizations pleasing and interactive for the user to understand different parts of the data in different views.

Bar graphs were chosen to represent the ranks of translated books and authors as they are best for comparing different groups. A line graph was chosen to visualize the number of books translated each year as it is best for showing trends. Packed bubble charts were chosen to easily compare the popularity of genres and languages to the rest based on differences in sizes. To give a variety of information in one view, we applied various filters in each section.

**Tableau Presentation and Analysis**

An interactive visualization using the LTI Korea data was created and designed using Tableau Desktop (see captured image below). This work is published on Tableau Public at [https://public.tableau.com/profile/publish/TransKoreanLit/Master#!/publish-confirm](https://public.tableau.com/profile/publish/TransKoreanLit/Master#!/publish-confirm). Interactivity of the visualization allows the user to view selected information in detail. When the user hovers over a section, such as a bar in a bar chart, a point on a trend line, or a section in a pie chart, a tooltip with detailed information appears and the associated graphs may change.
## Translated Korean Literature

### Literature Rank

Based on Literature Translation Institute of Korea’s bibliographic catalog, a total of 2,171 Korean literature has been translated from 1922 to 2015. We defined literature as books that fall into the category of fiction, poem, drama scripts, poem, and diary/journal. We excluded non-literary materials such as biographies, self-improvement books, children’s literature, and books with multiple or unknown authors. Below is the visualization of the data. You can filter by genre, language, or rank.

<table>
<thead>
<tr>
<th>Genre</th>
<th>Languages</th>
<th>Year Translated</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>(All)</td>
<td>(All)</td>
<td>(All)</td>
<td>1</td>
</tr>
</tbody>
</table>

1. 
2. 
3. 
4. 
5. 
6. 
7. 
8. 
9. 
10. 
11. 
12. 
13. 
14. 
15. 

![Graph showing the number of books translated over the years with a peak in 2015. The average years between the literature’s date of publication and translation is 16.87.]

---

39
Genre

There is a total of 8 genres.
* Classic refers to Korean literatures that have been written before mid-19th century.

Language

A total of X languages that have been translated into.

Authors Rank

There is a total of 565 authors that had their book(s) translated until 2015.
Literature Rank

In the Literature Rank section, users can see the rank of translated Korean literature. As the bar graph shows, *Please Take Care of My Mother* (엄마를 부탁해) by Kyung-Sook Shin is the most translated piece of literature. If the user hovers over a bar, a tooltip appears with information on the author, publication year, number of languages it has been translated into, and number of times it has been published. In addition, a book's cover image, year line graph, and translated languages section change to reflect the book's information. For example, when the user hovers over the *Please Take Care of Mom* bar, the book image on the right updates to its book cover and, under it, the translated languages section updates according to the relevant data; the year line graph also changes accordingly. When *Please Take Care of Mom* is selected, it shows 15 different years and the number of published books, representing the publication trend of the translated book. When users change the slide, they can view all the ranks. The ranks can also be re-sorted by language and genre at the user's discretion.

*Please Take Care of My Mother* was translated into 20 different languages and published 24 times. This book is also the most recently published book among the top 20 translated Korean books. Despite having the most translated versions of a book, the author Shin ranked fourth on the list of the most translated authors. One can hypothesize that the topic of the literature is the foremost important factor of its success across different cultures. The main theme of Shin's book, motherhood, is understood and empathized universally. It would be interesting to determine if this is evident across all top books. It would also be important to see the use of the literature outside of classrooms since the primary market for translated books are for course materials (Galasso and Park, 2017).

Locally renowned literature published after 2010 is not on the top 35 list, which comprises of 45 books. A few examples are *28* by Jeong Yu-jeong (published in 2013) and *Eungyo* (은교) by Park Bum Shin. One possible reason for the lack of post 2010 literature in the top list is the long length of time needed for translation. According to LTI Korea’s data, the average time for a book to be translated was 16.86 years (excluding the Korean pre-modern classics published before the New Literature era in 1894).

Year

The line graph shows an upward trend for translated books published between 1922 and 2015. The graph represents the fact that Korean literature was largely unknown to the world until the 1980s (Cerralbo, 2011). After the 1980s, the outside world showed an increased curiosity, and the government strongly pushed for the translation of Korean literature (Cerralbo, 2011). Thus, the path to translation was not always the norm. There was a major drop in the publication of translated books from 2005 to 2008, dropping from
130 published books in 2005 to 69 in 2008. Other factors affected the publication or translation market during that time, such as the 2007–2008 global financial crisis. The crisis is assumed to have affected the entire world, causing the book publishing market to diminish and thereby affecting the translation and publication of Korean literature. In 2009, the number of translated books began to increase again, and the total number of annual publications increased by 58 between 2008 and 2011. The rise could be linked to the new Korean Wave, *Hallyu 2.0*, that began in 2008 (Jin, 2012). Another point to note is the difference between the number of translated version of the top-rated books and all remaining books. The average number of translated versions for the top 20 books (0.07% of the total) is 15, compared to one for the rest of the books (98.3%). This shows that the market for translated Korean books is not evenly distributed. It is heavily concentrated on a few books published between 1979 and 2008.

Genre and Language

The pie charts in the middle show the distribution of genres and languages. English is the most translated language (22%), and fiction is the most translated genre (67%). It is not surprising that English is the most translated language since it is a global language that is being used in more than 80 countries (CIA, 2017). Chinese (16.48%) and Japanese (16.30%) followed, which was also predicted as they are neighboring countries to Korea with cultural similarities. An examination of the translated languages rank could reveal undiscovered relationships between Korea and other countries. Fiction (67.64%) as the most popular genre is not surprising since most of Korean literature is fiction. As for poetry translations, it is the second largest genre at 23.67%. Korean poetry has a long history since the Chosŏn dynasty, which could be the reason for its large proportion. However, it would be interesting to get more insights by comparing other cultures’ literature.

Authors Rank

The last section of the visualization is the rank of the authors. The top 30 authors who published the most translated versions of their work are, unsurprisingly, all authors who are renowned in Korea. The authors are ranked by the number of times their books have been translated and published. Yi Mun-yol is at the top; his books were published a total of 64 times. The bars next to the author’s name give information on the books published. The segmented sections within a bar indicate one piece of literature. For example, Yi Mun-yŏl has 17 pieces of literature (segmented sections) that were translated and published into different languages. *Our Twisted Hero* (*우리들의 일그러진 영웅*) was published 17 times and translated 13 times. The shades of a bar represent the number of pieces of literature. The more literature, the more teal; the less literature, the more yellow. For example, Ko Un has
28 pieces of literature, so the bar shows more teal, but Oh Jung hee only has 10 pieces of literature, so her bar lacks teal and instead is more yellow.

**Conclusion**

As we are in the field of librarianship, we sought to use bibliographic metadata for a visualization using the most current application (i.e., Tableau) so that anyone could understand the data in easier and more meaningful ways. At the same time, the goal of this mini-study was to take full advantage of Tableau, which is increasingly popular and highly used in data visualization. We hope our study offers some thoughts for future research topics for libraries and librarians to explore and utilize bibliographic metadata, which we have long invested in, maintained, and owned across libraries.

We not only learned the importance of the data source, but also realized that—even for a smaller scale study like ours—we have to rely heavily on good quality data. Without such quality, the research could be full of fundamental flaws. In addition, the data cleaning process could be an overwhelming job, further putting at risk the subsequent step of data analysis or visualization.

The most challenging issue for our study was securing the data set using open sources so that we could minimize the expense that might be incurred for any project like this. For example, OCLC World Cat records are freely available public records, but crawling through the bibliographic metadata from the current platform is limited in terms of enabling any public to obtain and use the data freely and comprehensively. We spent quite a lot of time figuring out possible approaches through OCLC, but ultimately none were feasible without going through their fee-based services in order to obtain the information we sought. Furthermore, as librarians, we lacked very specific skills and knowledge dealing with technology involving data crawling. Even with help from experts in those areas, we spent some time concluding that the data were only available through OCLC’s fee-based services due to the database’s complexity and large scale. It would be our wish for OCLC to provide the bibliographic data with a pre-developed tool enabling any researcher or member of the public to use the data set more easily.

Still much remains to be discovered if more comprehensive bibliographic data from OCLC could be easily accessible for future research. For example, the current study could be expanded to world literature and translated works. It would be fascinating to understand the relationship between world literature and translated works through languages, publication dates and countries, genres, and locations of holding libraries, as such an approach would provide a much bigger scale for visualizing the whole picture. Such efforts would highlight some interesting trends or phenomena in translation studies.

Using Tableau’s technology for data uploading and its convenience, it would be more meaningful and useful to update the translation data regularly if LTI Korea could provide them on a regular basis. Our study ends with the most current data (as of early 2016), but
the translated works keep growing every day, with more and more books being published in world languages. Capturing the most recent picture of translated literature today would be highly desirable.

This study has many limitations in terms of data set, analysis, visual presentation, and comparison with other Tableau presentations. Nonetheless, we hope our efforts have provided the opportunity to open up the discussion on potential use of bibliographic data for future related research involving cataloging metadata.

References


Managing Changes in Collection Development

Xiaohong Chen
Indiana University

Abstract

This article explores issues and strategies in collection management through the landscape of change in the academic library setting from the aspects of technology, the nature of the collection, the role of the librarian, and user services, with an emphasis on East Asian collection development and management. To cope with the new challenges East Asian librarians must develop collection tools, undergo training, collaborate, and learn how to communicate effectively.

Introduction

Change is the theme of the 21st century. New technology has widely changed the pattern of economic activities, social communications, and human daily life. Library services are not an exception. While academic libraries still uphold the fundamental mission to provide best information services for teaching, learning, research and community support, changes in many aspects of the academic library have challenged library professionals to cope with the new issues that have risen with changes. Moran and Brightman (2001, p.11) defined managing changes as “the process of continually renewing the organization’s direction, structure, and capabilities to serve the ever-changing needs of external and internal customers.” Collection management, the core component and the foundation of library information services, particularly requires attention as new challenges and issues arise. Managing changes means working with changing issues of technology, material and people to set up appropriate strategies, structures, and solutions to cope with changes. Borin and Yi indicates that the changing landscape of collections makes it even more important to understand our work (2011, p.120). By identifying new issues and finding solutions, collection managers can strategically handle collection activities and services in current library circumstances.

Change can bring to organizations and customers profound benefits. Change also often conflicts with organizational traditions and poses challenges to organizations and customers. In this article, the observation and analysis of changes will base on general phenomena in the academic library, with a particular emphasis on East Asian library collection management. In the field of East Asian collection management, like the mainstream library collections, the change can be mainly seen from these aspects: 1) the change of the nature of collections, from past physical only collections to present hybrid print and digital collections. For instance, in recent years, the components of the East Asian
Collection at Indiana University Libraries have expanded from the traditional physical items, such as monographs, periodicals, and microfilms to a hybrid print and electronic collection, including e-books, e-journals, electronic databases, audio-visual media, open access network and links. According to the 2016 statistical data, up to June 30, 2016, there were 2205 e-books, 8276 e-serials and 30 e-databases included in the total 338,345 CJK material held by the IU East Asian Collection (CEALS, 2016). Due to the change, collection librarians face more challenges in managing collection development, and users also face technology challenges for access and use of collection.

2) The change of the role of collection librarians. In the hybrid collection environment, a collection manager not only needs to manage a local in-house collection, but also needs to manage remote web collections through networks. The role of collection managers shifts from being a local collection builder to being a knowledge manager, an aggregator who works with a broad range of cooperators in and out of the library (Brin, Groen, & Thorin, 1999). The changing role requires collection librarians to enhance their capabilities, update knowledge and work skills, and use new strategies to accomplish collection management duties and be successful in playing the new role. East Asian librarians must work with local library administrators, patrons, and oversea vendors that require them to be familiar with both local policies and international operations, and to have multiple language skills.

3) The change of increasingly diverse user populations. In the past three decades, along with the change of the global landscape in political, economic and education policies, especially with the radical changes that have occurred in China, the community using the East Asian collections has undergone significant demographic changes. For Chinese collections, while they continue to be a resource center for the traditional East Asian studies programs at their campuses, the user population has greatly changed due to the increase in number of foreign Chinese undergraduate students and Chinese immigrants, the increase in number of Chinese language programs, and the increase in broad interest in China studies among American scholars (Zhou 2006, Shen 2009). The diverse user groups include faculty members and students who have Chinese language skills to use Chinese material for teaching, learning and research, American Chinese Scholars, American Chinese students, overseas Chinese students, local Chinese residents, Chinese family visitors, and local cultural organizations.

4) The change in acquisitions activities. On the one hand, new technology changes the methods of publication, communication, and acquisitions. Global network acquisitions and email orders shorten time and distance between North America and East Asian countries for oversea purchases. On the other hand, overload of information and publications from various sources come into conflict with tighter library budgets. According to a comparison of official statistics between 1988 and 2007 in Chinese publications, the 2007 total number of publications in China was 3.76 times of that in 1988. Meanwhile, the price of Chinese publications has increased from 20 times to 200 times. Despite the publication explosion and rising prices, the budget for Chinese collections in most American libraries has not increased
(Shen 2009, p.7). The situation has challenged East Asian collection librarians to work with vendors and consortiums in the business world and to make the best decisions for selection and collection, based on the organization’s mission, value, and users’ needs.

Based on the changes described above, the following research questions arise: What are the issues under the surface of changes in collection management, and what is the most effective way for collection managers to handle the issues? What does managing changes mean? Focusing on these questions, this paper will present three major issues that have been discussed by management professionals: 1) pressures caused by the change of the nature of collection; 2) the new roles of collection librarians; and 3) academic engagement in a business world.

**Issue #1: Pressures Caused by the Change of the Nature of Collection**

New technology has changed the traditional methods of publication and scholarly communication. As a result, library collections have changed from traditional physical collections to print and digital hybrid collections. In a broad discussion on managing the change, researchers mostly focus on how to manage hybrid collections, and how to make the new digital collections accessible (Gessesse 2000; Horava 2010; Branin et al. 1999; Tam & Robertson, 2002). Gessesse (2000) notes that traditionally, library materials were identified by reviewing sources, publishers’ advertisements, and approval plans. Now identifying internet resources presents more challenges because of rapid changes in web resources and lack of bibliographic control. There are also many new types of materials that are acquired in addition to traditional publications that are mainly electronic resources. Gessesse points out one of the major problems in the structural change for the collection is that collection librarians need to have an appropriate choice of what will be the chief characteristic for the information object, print or digital or both. Electronic resources are easy and fast to access, but they are at the same time tied to the compatibility of hardware and software, stability of resources and web services, and user instruction for e-access. Branin et al. (1999, p. 26) also points out that the introduction of digital resources in research libraries led to a good deal of conflict between the old and new formats.

The East Asian collection has encountered similar challenges with the increase of e-resources and significant changes in the nature of collections. For example, the Library of Congress has developed e-resources including Chinese maps, Chinese Rare Book Repository, digitized manuscripts, and online finding aids. Since 2004, the LC Asian Division has acquired 27 online databases on Chinese studies, most of which are subscription based, such as Duxiu, China Academic Journal, Century Journals Social Sciences, China Reference Works Online, Chinese newspaper databases, and Chinese Classics Databases. (Song 2011). In the new circumstances, East Asian Librarians need to adopt new ways of thinking in acquisitions and making decisions. As is true in collecting films, East Asian librarians should ask themselves if they should purchase a physical item DVD or if they should purchase an online streaming
film. Many librarians may still doubt if a virtual library can be reliable in terms of sustainability and accessibility.

Beyond the surface changes in the material formats in the collection, a further concern has to do with the people affected by the change of collection. Whittington (1992) indicates that the crucial issue is the way managers learn to balance the often-contradictory implications of change. Bamford & Forrester (2003, p. 557) interprets such a perspective to mean that “all of this is to be balanced within the ever-increasing pressures put on all levels within organizations.” In the context of library collections, changes in structure and format have caused a more complex workload for collection librarians (Branin, 1999; Tam, 2002; Horava, 2010). Another concern discussed by Demas & Miller (2012) is that in the new age of diverse collections supported by the growing corpus of digital surrogates, academic libraries will join networks of shared responsibility for storage and access to print and digital content. However, many library staff have a strong emotional investment in the physical collections and may be reluctant to participate in shared print programs. Change in the nature of collections has also caused user stress for access and use of collection. Traditional users and older users may have mental resistance to adapting to new use patterns and might feel anxious or upset when they encounter difficulty in discovering online information and using unfamiliar and continuously updated technology and application programs to access e-resources in the collection. These phenomena, which may also exist in other types of library environment, such as the public library, reveal the symptom of “personal immunity to change,” as described by Kegan and Lahey (2001, p. 51). Kegan and Lahey explain that many people are unwittingly applying productive energy toward a hidden competing commitment. A manager, like a psychologist, must go through the process of helping their employees overcome their immunity to change, to guide people to understand the necessity of change (2001, p. 52). Thus, adopting appropriate strategies to release the pressures deriving from such a change is equally important for both collection librarians and collection users. For library collection management, managing change means not only dealing with the change of the nature of collections but also dealing with how to work with people in a changing environment.

Strategies for reducing psychological techno-stress among librarians and users have been summarized by library collection professionals in the following ways: 1) adopting updated management tools for helping collection librarians effectively manage complex print and digital collections. Wilson advocates using an integrated management system consisting of several subsystems to manage in-house collections, electronic collections, and user support (Wilson, p. 1998). For developing and selecting useful electronic resources, Gessesse advises using appropriate internet resource tools such as Internet Compendium and WWW virtual Library to identify internet resources (Gessesse, 2000). Using appropriate strategies can help to reduce mental resistance caused by changing collection content, format and use patterns, as well as to make the full collection effectively used. The American-based OCLC bibliographic information services include over four thousand OCLC member
institutions in the Asia Pacific region. In East Asian collection practices, the OCLC electronic database provides great convenience for CJK collection acquisitions in terms of online bibliographic control to obtain title information, format, and title holding status. In addition, the integrated library system allows library collection acquisitions to share order information with oversea vendors.

2) For convenient access to the digital and print collection, it is necessary to improve the design of the collection webpages in terms of clearance and convergence, to compile a subject guide of electronic resources, and to select, organize, and clearly display useful information resources and make them accessible to the user (Tam et al. 2002). Georgas (2014, p. 528) studies the information-seeking behavior of undergraduate students within a research context and concludes that libraries must continue to work with vendors to improve electronic search tool resources with the user in mind and with a goal of providing increased findability of contents, flexibility, search algorithms, relevance of results and ease of use. For the East Asian Studies collection, the OCLC Connexion module integrated CJK cataloging and CJK scripts (Shen, 2009, p.7) into catalog records. This development is good for East Asian librarians, acquisitions specialists, catalogers and library CJK material users faster and more accurate grasp bibliographic information, since reading native language scripts is much easier, faster and more understandable than reading Romanized characters for either East Asian library staff or library patrons. The Indiana University Libraries, where I have been working for several years as East Asian Acquisitions Specialist, recently completed the project of adding Chinese scripts to all Chinese material bibliographic records in our workflow MARC database, including older Chinese material in the collection.

3) Provide training and support services. To assist users in accessing and using electronic resources, collection librarians need to provide proactive support, technical training, and information literacy instruction to users (Tam et al., 2002, p. 371). Collection librarians may need to set up user groups with goals geared to the particular needs of different users. For undergraduate students, sessions may focus on information literacy and a research tool guide. A study on citations from papers written by first year undergraduates revealed that the students used more books, more types of sources, and more overall sources when a librarian provided instruction (Cooke & Rosenthal, 2011). For faculty, instruction sessions may need to present new media and database services with the goal of increasing awareness of the e-collection. For instance, the IU Libraries East Asian Collection website provides users online Lib-guide, resources suggestions, and bibliography tools. The Ease Asian librarian periodically announces newly acquired e-resources, e-databases for trial via the mail-list distribution to the East Asian users group. We also have workshops during academic year for campus-wide faculty and students to guide them in searching and using both physical and electronic resources in the collection. In March 2017, the East Asian Librarian at IU-Bloomington campus held a workshop with two sessions on the topic “Researching China: Electronic Access to Primary Sources and More,” introducing to IU users many unfamiliar Chinese electronic databases, including electronic journals, dissertations,
newspapers, China digital library, Chinese Ancient Text (CHANT), and Adam Matthew Digital Collection of China studies, etc. Users can easily access these full text databases via the aggregated web links on the IU library website for the East Asian Studies.

![Figure: A web image of the East Asian Studies Collection, Indiana University Libraries](image)

**Chinese Studies -- Internet Resources**

*Dictionaries and Encyclopedias*

- Chinese Characters Dictionary Web
- China Biographical Database (中国近代人物传记数据库)
- CCL (Center for Chinese Linguistics PKU) 资料库
- Dict.cn（词典）
- English Chinese Glossary of Computer and Network Communications Terms (英汉计算机及网络通信术语库)
- Thesaurus Sinicae (TLS) 汉语文典

**Issue # 2: New Roles of Collection Librarians**

Technology changes the nature of a library collection, the use pattern of the collection, and the service components of the collection. As a result, it also changes the role of collection managers. A survey conducted by Dorner Daniel (2004) aimed at examining quantified changes of the role of collection managers in the overall structure of information services arising from the arrival of digital information. The findings showed that over the five years preceding his survey, collection managers’ roles had been evolving, and the levels of responsibility and the time spent on activities related to digital resources had increased for most collection managers. The roles of collection managers have blurred, in that they continue to perform functions related to traditional print-based resources while performing new functions related to digital resources and spending more time on consortia work.

Once collection managers begin to participate in the collection of digital materials, their working scope expands greatly in terms of environment, relationships, and duties. Beyond the traditional roles of controlling bibliography, selecting and deselecting items, providing subject reference assistance, and collecting and organizing materials inside the library, collection managers now need to play multiple roles. They may work as a
coordinator to collaborate with technology departments to maintain webpages and digital collections. They may also work as an aggregator to network the information in the subject area and exchange collection information with other institutions, through participating in the centralized online database and cooperative collection programs, as an instructor to participate in library online learning and workshops for both local and remote users. They also play a role as a trader negotiating bargain prices with vendors worldwide by emails and via buying trips. Thus, today, collection managers must juggle multiple tasks and become proficient at many skills. They not only manage the information knowledge through physical and digital collections, but also manage the environment of the collection with multiple work units and organizations, local and global, on-site and outside. As Branin et al. (1999) describes the present situation, a traditional collection development perspective, which emphasized only acquisitions, selections, and building collections, has moved toward a new vision of collection management and a much broader range of collection performance including collection policy development, material budget allocation, selection, collection analysis, collection use and user studies, preservation, and cooperative collection development.

The fascinating new roles for collection managers bring with them the issue of how these new roles can fulfill successfully considering the new duties and demands on the time of collection managers. Some suggestions for how to deal with this issue have been put forth. Learning new knowledge and taking training programs for new technology are one way to transfer from traditional roles to the new roles. Forte et al. (2002, p. 305), in a case study examining collection manager training at UC-Santa Barbara, indicate certain ways to refresh collection managers, such as creating a standing committee to supervise training, updating the collection managers’ manual, and repeating the orientation program for all collection managers on a recurring schedule. Coe & Consoli (1994) address how to train collection librarians in an environment that is transitioning from the bibliographer model, where selection and collection management is the sole responsibility, to an environment where librarians from other library departments, and with other responsibilities. East Asian Library organizations have taken actions to help East Asian librarians cope with the challenges of change in technology and collections through training programs. East Asian library committees have held a few training workshops for librarians. The events include 2001 Korean Materials workshop, 2002 Junior Japanese Studies Librarian Professional Training Seminar sponsored by the North American Coordinating Council on Japanese Library Resources (NCC). In 2004, sponsored by the Henry Luce Foundation and in cooperation with the committee on Chinese Materials of the Council for East Asian Libraries, the University Library System at the University of Pittsburgh organized the Luce Summer Institute for East Asian Librarianship: China Focus. The three-week summer program provided training for Chinese studies librarians and library managers focused on the latest developments in Chinese language software and application of Unicode, digitization of Chinese language materials and technical processing of Chinese language material, and
current awareness of access to electronic resources for East Asian studies in general and Chinese studies in particular. Topics also included book trade and vendor relationships, acquisitions, astute management of limited budgets, faculty liaison, research methodology and use of East Asian studies resources (Shen & Wei, 2003). The goal was to update librarians to help them manage their many roles, particularly collection development, in the changing world.

Another valuable idea offered by management expertise that may help librarians cope with the issues facing them is to accomplish goals through collaboration. Bamford et al. (2003, p. 557) indicate that the role of managers is not to plan or implement change, but to create an organizational climate that encourages experimentation and risk-taking, and to develop a workforce that will take responsibility for identifying the need for change and implementing change. With this principle, collection managers should direct and structure the work of employees under their supervision to work together to get things done, rather than doing everything themselves. Gosling and Mintzberg (2003, p. 56) suggest that “one obvious answer is: about collaboration, about getting things done cooperatively with other people-in negotiations, for example, where a manager cannot act alone.” They elaborate that to have a collaborative mind-set means getting away from the currently popular heroic style of managing and moving toward a more engaging style. They foster collaboration among others, and get things done, but they do not do things alone. “They are on the top of a network, looking down on it, to be in a collaborative mind-set, to be inside, involved, to manage throughout, and to get management beyond managers, to distribute it so that responsibility flows naturally to whoever can take the initiative and pull things together” (Gosling & Mintzberg, 2003, p. 60).

In such a collaboratively oriented management style, collection managers can work with the IT and technical services departments in the library to manage online collection activities. They can also work with other library institutions and consortia to share information resources and build collaborative collections. Dorner’s survey findings show that the vast majority of collection managers are liaising with computing staff to resolve technology issues related to digital resources, rather than doing this work on their own (2004, p. 267). Gessesse suggests that collection librarians may seek to develop e-resources through resource sharing and cooperative collection in a networked e-environment (Gessesse, 2000). Shen points out that many national and regional consortia have been formed to facilitate resource sharing. Chinese collections usually join their home institutions in regional library consortia. The consortia promote scholarly communication, interlibrary loan, shared electronic resources, cooperative collection development, digital libraries, and help for librarians to optimize budgetary and space constrains. An example is a collaborative collection development arranged between the University of North Carolina and Duke University for Chinese and Japanese materials (Shen, 2009). Troost shares an idea for learning from others when East Asian librarians, posted queries to the listserv, Eastlib, or email other librarians with a difficult reference question. “I routinely pick up other people’s
brains because they will say something I have not thought about and I will also gain from their reactions when I share my ideas.” (Troost, 2009, p.2). In an incident in my recent experience, the East Asian librarian at my home institution, when deciding whether to purchase a sixty volume Japanese book series, considering the cost, chose to purchase only the index volume, on the rationale that users with a research need, by looking at the index, can request a specific volume they need via interlibrary loan. In the digital age, sharing resources globally through the institutional collaboration has become feasible and significant. From the IU Libraries website for East Asian Studies, users are able to explore numerous open access digital resources that are created and maintained by other institutions for their teaching, learning and research needs.

The collaborative model is a way for collection managers to lessen the number of necessary tasks through collaborating with best partners to achieve effective collection development. In sum, dealing with the issue of new functions and new roles in collection practice, collection managers may overcome the challenges through seeking training and collaborative engagement.

**Issue #3: An Academic Engagement in a Business World**

Technology is changing the material world. Technology is also changing time spent on processing and collection activities through speedy communication and speedy delivery with high productivity. Today, collection librarians and acquisitions staff use the Internet to connect and contact worldwide vendors and publishers to acquire collection materials in a direct and fast way. Thus, the quantity of collection can be increased faster than ever. For instance, in the East Asian acquisitions processes, from placing orders with vendors in China, Japan and Korea to receiving ordered items, the time period has been generally shortened from over six months in the past to nowadays around two months. However, the increased prices of publications, especially for journals, have “literally eaten up limited collection budgets” (Branin, 1999, p. 26). One consequence is that monograph purchases have decreased (1999). Librarians must carefully manage funds in order to meet the demand for both print materials and digital materials (Gessesse, 2000).

Conflicts between budgetary constraints and economy in acquiring materials lead collection librarians to seek cost-effectiveness and efficiency in acquisitions. To deal with the issue, Tam et al. (2002) suggest that librarians can use consortial power to obtain better prices for information resources, apply the principles of e-commerce to the library work process, and adapt business system design methods to library management. In addition, they can find less expensive e-publishing alternatives to print resources, sharing resources through open access, organizing co-operative acquisitions of scholarly publications, and closely working with faculty to ensure best resources collected. Demas and Miller (2012) argue for cost-effective stewardship through collaborative collection management by which libraries will share responsibility for storage and access to print and digital content through the corpus of digital surrogates, rather than independently-owned collections. These
suggestions sound like smart business strategies. To engage in these activities in collection management, a key issue is how to strengthen communications between collection managers and other parties to get good deal in the business.

Clearly, effective communication is vital in managing change. Kotter (2007, p. 8) points out, “Without credible communication, and a lot of it, the hearts and minds of the troops are never captured.” Agee (2007) writes that collection librarians can find ways to obtain funds and negotiate prices through effective communicative engagements, stating that writing a successful grant application to purchase additional materials for the collection is an art. Building a good relationship with successful business organizations and wealthy individuals might lead to financial sponsorship or donation for the purchase of library materials. Negotiating with sellers is often necessary and usually worthwhile because the vendor often allows a discount for all orders and may also pay shipping costs for the materials. Establishing book exchange program through liaison with other libraries may also be financially feasible. Demas and Miller (2012) suggest that to manage collections collectively, a collection management plan should be written to articulate for both internal and external audiences policies, criteria, and guidelines, to secure support from library staff, administration, faculty, and outside stakeholders. A carefully managed plan and effective communication can increase cost-effectiveness and prevent complaints and criticisms. Thus, to seek successful cost-effective collections, collection librarians need to act as a middleman to articulate and negotiate with vendors, publishers, consortia, donors and collection users to obtain the best price and build the best-valued collection. These business measures have noticeably played in East Asian collection activities in recent years. We chose book vendors to work with for the reasons of material they can obtain for us, fast and quality services they can provide, and the discounts they can offer. We negotiate with them and build good relationships with them through understanding their cultures and through effective communication. As a result, we have obtained more valuable, less expensive material to add to the East Asian Studies Collection.

**Conclusion**

In general, in the changing world of collection management, the structure of collection and access to collections are changing, the roles of collection managers are changing, and acquisitions activities are changing. Thus, new issues arise associated with these changes, such as the nature of collection, techno-stress among collection librarians and users, the complex duties and overwhelming workload of collection librarians, and budget constraints in acquisitions. These issues challenge collection management. To deal with the challenge and to achieve effective collection development, we need to adopt effective strategies to manage changes, such as developing useful tools, holding appropriate training programs, engaging collaborations and effective communications. Managing change in collection development means to us not only managing changes in technology, budget and material, but also managing changes of people, book vendors, collaborative partners, team members,
and the users of collection. This ideal is applicable for any types of library environments, including East Asian Libraries, because they face similar changes and issues.

Bibliography References


Korean Romanization: Is It Finally Time for the Library of Congress to Stop Promoting McCune-Reischauer and Adopt the Revised Romanization Scheme?

Chris Doll
University of Dubuque

According to the Korean Tourism Organization (KTO homepage) in 2012, 11.1 million foreign tourists visited South Korea, making it the 20th most visited country in the world. In addition to the fact that as of 2013, 1.5 million foreigners also lived in Korea (Yonghap News, 2013) it is no secret that the Korean government has done all it can to Romanize and standardize hangeul (Korean language). With hallyu (Korean wave) taking off, the Korean Government is riding this wave by promoting Korea through the use of English. In order to do this, they have tried to standardize place names, people’s names and local food. This has been a laborious project because hangeul is not easy to Romanize. As Horace G. Underwood, former director of the Yonsei University Library, pointed out back in 1972: “The scale of vowel sounds does not really fit any Roman system, and the basic structure of the consonants, with aspirate-non aspirate, and without voiced-unvoiced phonemes, cannot readily be represented in an alphabet that recognizes voiced-unvoiced but not aspirate-non-aspirate distinctions.”

Because of the aforementioned problems, Korea now has two main Romanization systems; the Library of Congress (LC) promotes the phonetically based McCune-Reischauer (MR) to Romanize Korean words, whereas the Korean Government supports its own form of Romanization, otherwise known as the Revised Romanization (RR) scheme1. As a result, Korean words are often Romanized using various methods. Most of the world and general knowledge books about Korea generally follow the RR method while Korean language learners and academic journals tend to still follow MR. This is a problem because information becomes discombobulated. In order to do a thorough search on a topic one needs to not only know both forms of Korean Romanization, but also advanced search techniques. This causes information retrieval and storage about Korea to be both time consuming and inefficient. The table below shows the MR2 and RR3 scheme side by side with the differences in bold. As you can see, the majority of the variances are due to the reliance on diacritics when using McCune-Reischauer (as you can see with the vowels ㅏ, ㅑ, ㅓ, ㅕ, ㅗ, and the consonants

---

1 This paper will use the RR method when Romanizing Korean words unless otherwise noted.
2 To view the more complete Library of Congress version of the McCune-Reischauer system please refer to http://www.loc.gov/catdir/cpso/romanization/korean.pdf
3 To see a complete RR table please refer to http://www.korean.go.kr/front_eng/roman/roman_01.do
Some of the secondary differences can be attributed to pronunciation (as is the case with ㅓ, ㅗ, ㅘ, ㅝ).

<table>
<thead>
<tr>
<th>Vowels</th>
<th>McCune-Reischauer</th>
<th>Revised Romanization</th>
<th>Consonants</th>
<th>McCune-Reischauer</th>
<th>Revised Romanization</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Initial</td>
<td>Final</td>
<td>Initial</td>
</tr>
<tr>
<td>a</td>
<td>a</td>
<td>k</td>
<td>k</td>
<td>g</td>
<td>k</td>
</tr>
<tr>
<td>ae</td>
<td>ae</td>
<td>kk</td>
<td>k</td>
<td>kk</td>
<td>k</td>
</tr>
<tr>
<td>ya</td>
<td>ya</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
</tr>
<tr>
<td>yae</td>
<td>yae</td>
<td>t</td>
<td>t</td>
<td>d</td>
<td>t</td>
</tr>
<tr>
<td>ŏ</td>
<td>eo</td>
<td>tt</td>
<td>-</td>
<td>tt</td>
<td>-</td>
</tr>
<tr>
<td>e</td>
<td>e</td>
<td>r</td>
<td>l</td>
<td>r</td>
<td>l</td>
</tr>
<tr>
<td>ê after and</td>
<td>e</td>
<td>r</td>
<td>l</td>
<td>r</td>
<td>l</td>
</tr>
<tr>
<td>yŏ</td>
<td>yeo</td>
<td>m</td>
<td>m</td>
<td>m</td>
<td>m</td>
</tr>
<tr>
<td>ye</td>
<td>ye</td>
<td>p</td>
<td>p</td>
<td>b</td>
<td>p</td>
</tr>
<tr>
<td>o</td>
<td>o</td>
<td>pp</td>
<td>-</td>
<td>pp</td>
<td>-</td>
</tr>
<tr>
<td>wa</td>
<td>wa</td>
<td>s</td>
<td>t</td>
<td>s</td>
<td>t</td>
</tr>
<tr>
<td>wae</td>
<td>wae</td>
<td>ss</td>
<td>t</td>
<td>ss</td>
<td>t</td>
</tr>
<tr>
<td>oe</td>
<td>oe</td>
<td>-</td>
<td>ng</td>
<td>-</td>
<td>ng</td>
</tr>
<tr>
<td>yo</td>
<td>yo</td>
<td>ch</td>
<td>t</td>
<td>j</td>
<td>t</td>
</tr>
<tr>
<td>u</td>
<td>u</td>
<td>tch</td>
<td>-</td>
<td>jj</td>
<td>-</td>
</tr>
<tr>
<td>wŏ</td>
<td>wo</td>
<td>ch’</td>
<td>t</td>
<td>ch</td>
<td>t</td>
</tr>
<tr>
<td>we</td>
<td>we</td>
<td>k’</td>
<td>k</td>
<td>k</td>
<td>k</td>
</tr>
<tr>
<td>wi</td>
<td>wi</td>
<td>t’</td>
<td>t</td>
<td>t</td>
<td>t</td>
</tr>
<tr>
<td>yu</td>
<td>yu</td>
<td>p’</td>
<td>p</td>
<td>p</td>
<td>p</td>
</tr>
<tr>
<td>ŭ</td>
<td>eu</td>
<td>h</td>
<td>h</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ŭi</td>
<td>ui</td>
<td>i</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“In North America, the ALA-LC Romanization Tables are an established standard for library cataloging, but libraries elsewhere in the world are more likely to use the various ISO Romanization standards or a national standard. Often, different standards result in very different Romanized strings that may, at best, look strange (and, at worst, not be recognizable) to a user accustomed to what is done in another country. They can also wreak havoc with attempts to match records. And MARC 21, unlike UNIMARC, has no way of indicating in the bibliographic record which Romanization practice has been applied” (ALA,
This excerpt from the American Library Association (ALA) not only reiterates the problems multiple Romanization methods cause for users, librarians and catalogers, but it also shows that ALA and the Library of Congress often like to do things their own way. Why can’t ALA and the Library of Congress be like the rest of the world and adopt the national standard? This paper intends to question whether ALA and the Library of Congress should be so cavalier in their desire to continue the use of their sometimes outdated ALA-LC Romanization Table, in this case MR, and examine whether a shift away from MR would benefit users, librarians and catalogers.

An outright shift overnight is impractical, but more research should be done to see if a gradual shift is feasible and necessary. It appears as if the contributors to the most up-to-date ALA-LC Romanization Table were aware that this day would come. In their online PowerPoint that was used as an introduction to the revised guidelines for the 2009 edition of Korean Romanization they included a slide entitled “In The Future” in which they added: “If, in the future, a strong consensus of Library users form around another specific Romanization system or set of practices, Library of Congress would be most willing to revisit and reconsider these MCR Romanization and word division guidelines” (Y. Lee, 2009). In addition to this, there have been papers in the past that have also pointed out the confusion caused because of the two systems. The reason it should be examined again is because more people are becoming interested in studying and learning about Korea. In the past, in America, it was mainly scholars or Korean immigrants who were searching for information about Korea. Most Korean scholars are not only familiar with MR, but are also fluent in Korean and would easily overlook the problems that are prevalent on a basic level. Many of these scholars have gone on to work for ALA, the Library of Congress or are now in charge of the Korean Studies section of a library. Using hangeul or MR is second nature by now. But since we have such an influx of new people interested in Korea, wouldn’t it behoove us to ensure that information is as readily available as possible? Shouldn’t we remove as many roadblocks as possible? Shouldn’t we make sure that Korean Romanization is easy enough for non-Korean specialist librarians to properly help a patron in need of information about Korea? Studies should be done to see if information about Korea is easily attainable, or whether a convoluted Romanization scheme inhibits this search.

This paper will examine the benefits of a potential change from MR to RR. This will be done in four ways. First, this paper will look at what Romanization is, why having multiple systems causes confusion and errors, why Romanization methods are not archaic and why it’s still relevant to users, librarians and catalogers in modern time, especially when using Korean. Second, prove that MR is antiquated and inefficient. Third, show that RR is more effective than MR and is also not only a short term solution, but a long term answer. Lastly, show how a similar transition in the past has proven to be successful.
Romanization

Romanization is “the process of writing or transliterating a non-Latin character into a Latin character” (Tull 2003, 442). This is done for a variety of reasons. It can be for writing academic papers, explaining current events in newspapers, or it can be a simple restaurant review. The need for Romanization is clear; it is not practical for anybody to learn every single language. Instead there needs to be a common or standard way to transcribe something in a non-Latin language into a Latin or Roman alphabet in order for there to be a universal way of phonetically reading these words. Although there are many countries in the world that face this issue, especially in Asia, a lot of focus has been placed on China, Japan and Korea. In fact in 1957, Preliminary Rules and Manual for Cataloging Chinese, Japanese and Korean Materials was published following the need for standardization of Chinese, Japanese, and Korean (CJK) materials (Cheng, 2000). The strong need for this was due to more people wanting information on Asia, specifically China, Japan and Korea. More and more libraries were housing Asian works and standardization was needed. Although libraries and other information agencies recognize the need for standardization, it was not and continues to not be an easy process. This is because, “it is complex work to create a transliterating or Romanizing standard because the pronunciation and morphology of East Asian languages are extremely different from languages that use the Roman alphabet” (Kim, 2006, 58).

In an effort to standardize the way non-Latin script data is cataloged, since the 1980’s, LC has distributed MARC records that include non-Latin script data in bibliographic records and in references in authority records. Starting with Chinese, Japanese, and Korean, LC now includes support for many of the languages that use the Cyrillic, Greek, Hebrew, and Perso-Arabic scripts. To do this, LC has implemented automatic transliteration capabilities known as “Transliterator” (LC, 2011).

Brief history of Korean Romanization methods

The McCune-Reischauer system (MR) was devised by G. M. McCune and E. O. Reischauer in consultation with Korean linguists. It was first published through the Korean Branch of the Royal Asiatic Society in 1939. The MR system has been widely used for historical, literary, political, and military purposes and in the general presentation of Korean proper names in a Romanized form (Kim, 2006). As the original innovators of MR contend, MR is heralded for its: “scientific accuracy and practical simplicity” (McCune, 1939). Because it is still the Romanization system favored by LC, it is still the preferred form of Romanization in most academic journals, specifically Korean Studies and East Asian Studies journals.

Although MR is used by LC and is still widely used overseas from Korea, it is not the official Romanization system used by the Korean government. On July 4, 2001 the Ministry of Culture and Tourism introduced its new Romanization system (RR). All geographical names now follow this system without exception. In addition to this, as of February 28, 2002 all textbooks now use this system and as of December 31, 2005 all road signs were changed
over. One of the exceptions was previously used Romanization of personal and company names. This is fortunate for Samsung because under the new translation rules the company name would be changed to *Samseong*. This new system was needed in Korea because native Korean speakers, in particular, have more difficulty than others in understanding the MR system, because it was devised for non-native speakers (Kim, 2006, 59). In addition, as mentioned above, the breves and apostrophes caused confusion. The new government system eliminated these confusions.

*Complications due to dual Romanization systems*

Having multiple Romanization systems that are supported by various institutions can cause a bevy of problems. To illustrate the complications of having two Romanization systems, a simple search query was performed using common Korean words that were Romanized into RR and MR. The terms were input into a search engine (Google) and also into an online catalog (WorldCat). The first subject entered was the *Joseon* Dynasty (조선시대), which was the last of the Korean dynasties, lasting from 1392 until 1897. To begin, Joseon Dynasty (RR) was entered into Google and the first 20 retrievals were recorded. The same procedure was repeated, this time using Chosŏn Dynasty (MR). The results can be seen in Appendix A. Retrievals that appear on both lists have been put in bold. As you can see, only eight of the twenty retrievals were on both lists. In order to get a complete and full understanding of the *Joseon* Dynasty, you would need to not only know both the RR and the MR Romanization system, but you would have to do two separate searches or do an advanced search that includes the terms Joseon and Chosŏn. It is also interesting to note that only one of the links retrieved (the very last one in fact) has Chosŏn with a breve, which is the correct and official Romanization style of MR and the Library of Congress.

The same experiment was conducted again, this time into OCLC’s (Online Computer Library Center) WorldCat, the world’s largest library catalog. The difference between *Joseon* Dynasty and Chosŏn Dynasty when retrieving information is even more evident when using these terms in WorldCat (see Appendix B for the results). As you can see, this time there was only one similarity out of twenty. WorldCat is used by a variety of people, organizations and libraries to retrieve resources on a subject matter. If a user was familiar with only one of the Korean Romanization methods, then the user would severely handicap their potential to fully research a topic.

Once again, the experiment was repeated, this time using Seoul (RR) (서울) and Sŏul (MR), the capital of South Korea. Because this is the exact same place, it should be expected that all twenty retrievals are the same. However, the opposite proves to be true. From the information gathered in Appendix C, only one of the twenty retrievals in Google was on both lists. The list that was compiled using “Seoul”, which is the Korean Government’s way of Romanizing the capital city, has many more relevant searches in regards to the city itself. The list with “Sŏul” has less to do with the actual city and more to do with history or specific
venues. Once again, the difference between the two Romanization retrievals is magnified when examining the search results in WorldCat (Appendix D). WorldCat does not recognize diacritics. As a result, it thinks of Sŏul as a soul, like the soul of a human. This leads to zero matches from the list and if you were looking for information about capital of Korea and you typed in Sŏul you would not retrieve any relevant results.

This experiment shows that a single search query retrieves extremely variant search results depending on which Romanization method you use. This problem is compounded if one is searching using WorldCat, which is supposed to aid patrons who want to take advantage of services that a library provides. Because LC promotes a Korean Romanization system that uses diacritics, which is something the world’s largest library catalog does not recognize, a user can be left with very low precision and may not retrieve any pertinent information at all.

Why Romanization is still necessary

In 2008, James Agenbroad published a paper entitled Romanization is not Enough in which he argues that “for finding non-Roman script library materials, catalogs with Romanized access points alone are inadequate because they are un-familiar to those who seek these materials.” Although Romanization is not a simple task and it can cause information retrieval problems if there are multiple systems in use, it is necessary. The ALCTS Non-English Access Working Group on Romanization Report (ALA, 2009) and the Library of Congress’ Romanization Landscape (LC, 2011) do a great job of explaining why Romanization is still relevant, especially for libraries. This paper will highlight a few of those reasons and explain why they are specifically important in regards to Korean Romanization.

Agenbroad builds his argument around the premise that “in America many public and academic libraries acquire, organize, house, and provide resources that use non-Roman script for immigrants from and students of places that use these scripts.” But that’s just not true. As LC points out, “Romanization is primarily for LC staff and staff at other libraries without language expertise working in Circulation, Acquisitions, Serials check-in, Shelflisting, Shelving, and Reference. Romanization is also for systems that cannot use non-Latin forms, have support for only some scripts, or require Romanized fields for indexing and sorting purposes.” (LC 2011) Romanization is for patrons AND library staff who cannot read non-Roman script who work with library materials in these scripts for various purposes (acquisitions, ILL requests, storage retrieval requests, assembling bibliographies). Many public libraries collect material in a wide variety of languages to serve linguistically diverse user communities, but are unable to employ specialists in all these languages. Even large research libraries are unlikely to have staff in every department who is able to interpret all the scripts used in the material they need to process." (ALA, 2009) Although Korean Studies is becoming more prevalent, many libraries do not have a full time Korean Studies librarian so Romanization is needed to perform basic library staff duties if Korean resources are present at a library.
Romanization is also necessary in regards to technical services. Romanization allows the staff to easily reproduce in writing and pronounce unfamiliar words when communicating with patrons about recalled items or fines or in phone conversations with vendors. This might be otherwise impossible in Korean. Romanization also helps library patrons. Patrons often search library catalogs based on people and places mentioned in Roman-script newspapers and other publications. (ALA, 2009) Especially in Korean the original-script form cannot be easily determined. Having Romanized data indexing in the catalog makes it easier to search. This is another reason why the Library of Congress should consider adopting RR over MR. All of the Korean newspapers that publish in English and many Roman-Script newspapers use RR. In addition to this, the world recognizes the RR spelling of cities not MR (as evidenced in the prior research of Seoul vs. Sŏul). People who want more information about the Korean capital and have read the word Seoul in a newspaper or see it printed on a CNN telecast would likely not know how to do a search using 서울 (Seoul) (nor would they know how to read the corresponding results) so Romanization is necessary. Also they are probably not familiar with the variant spelling Sŏul either.

Another advantage is that Romanization provides collocation when the same word can be written in different ways in the original script. Many historical Korean texts are written in both Korean and Chinese. For example Hanguksa ("History of Korea") can be written 韓國史 in Chinese and 한국사 in Korean. A search for the Romanized form retrieves all these variants. Romanization provides collocation when the same word can be written in different ways in the original script. Many systems are not sophisticated enough to treat these original-script forms as equivalent in their indexing but Romanization allows this (ALA, 2009).

Another reason that Romanization is still necessary is in regards to sorting. Doing a browse search for Romanized text gives an alphabetical list. In CJK sorting by code point is inadequate and is not easy to alphabetize. Romanization allows Asian scripts to be listed alphabetically which is universally known and recognized. This type of sorting also helps librarians who themselves are searching or the users they are assisting in finding Korean materials. In fact, according to a survey done by Jeong (2009), 79% of librarians searching for Korean Materials in an OPAC prefer Romanization over Korean scripts.

Lastly, Romanization is helpful because it provides pronunciation-based access points. This is beneficial for users who know the basics of a language but are not fully proficient in the original script. This is especially useful for undergraduates, beginners, or researchers who are not yet experts in these languages (ALA, 2009). In regards to Korea, since 2010 Korean as a second language has risen 62% (Chosun Ilbo, 2015). In addition to this, the number of Korean Studies departments and majors is on the rise. More Korean resources are needed for non-native speakers and Romanization certainly adds pronunciation based access points.
McCune-Reischauer is Antiquated and Inefficient

One of the most appealing aspects of the MR model of Romanization is also its biggest roadblock. The MR system aims at representing the pronunciation and not the spelling of Korean words. This is good if you are learning the language but it can be problematic if you are trying to retrieve information using Romanized Korean words. It is not a method of transcribing Korea script letter by letter into the Latin alphabet, but instead, it takes full account of euphonic changes in the pronunciation of Korean letters (Royal Asiatic Society, 1961). As a result of this, the MR system is full of breves and apostrophes. However, quite often the breve and apostrophe are omitted when writing. The breve is not a standard key on a keyboard and many non-native Korean speakers do not feel the need to add apostrophes (National Academy of the Korean Language, 2001). Research by Jeong (1998) and Kim (2006) also conclude that many users don’t use diacritics. Omitting diacritics can cause an extreme amount of confusion. Take for example Shinch’ŏn (신천) and Shinch’on (신촌). Not only are both of these places in Seoul, but at one time they were both on subway line 2 (the circle line). They are on opposite ends from each other. If you leave out the breves and apostrophes, they are both spelt Shinchon. If you looked up information for a restaurant and the information told you to go to Shinchon Station exit 4, you could have ended up going to the wrong subway station on the wrong side of Seoul. This caused a lot of confusion and inevitably Shinch’ŏn (신천) station was renamed Jamsilsaenae. 

MR’s reliance on diacritics also causes a problem when searching for Korean titles when using an ILS (Integrated Library System) which is more often referred to as a library’s online catalog. Some of the more popular ILS that are being used by academic and public libraries that were included in ALA’s Library System Report 2017 in no particular order include: Innovative Interfaces Polaris and Sierra, Apollo, Horizon, Koha, WorldShare, and for good measure I have included EBSCOhost which is mainly used for academic journals. I found a library using each of these ILS platforms. Polaris-Davenport Public Library, Sierra-Santa Clara University, Apollo-Georgetown Public Library, Horizon-Jessamine Public Library and Stanford University, Koha-Carnegie Stout Public Library, WorldShare-University of Dubuque, and EBSCOhost-University of Dubuque. In each of these ILS I entered the word Sŏul. In Polaris, Apollo, the standard version of Horizon that Jessamine Public Library uses, Koha, WorldShare and EBSCOhost the same occurrence that happened in WorldCat happened again, the diacritic was dropped and all of the results were focused on a human soul, or soul music, etc. and not the capital of South Korea. Sierra is an upscale and more expensive version of Polaris and was able to recognize that Sŏul was the capital of South Korea and half of the results that were retrieved were about souls, but half of the results were about the capital of Korea. However, the retrieved results mainly included the RR version of “Seoul” and not the MR Sŏul. In addition to this, SirsiDynix, the producer of Horizon, worked with Stanford University to support Asian and other multi-byte character sets. As a result, the Stanford ILS was also able to retrieve results centered on Sŏul; however,
just like with Sierra, half of the retrievals revolved around other types of souls. So in conclusion, the basic ILS that the majority of public and academic libraries possess will not recognize Korean diacritics. Advanced and thus more expensive ILS will recognize Korean diacritics, but are far from perfect and the retrievals are not as accurate as they are if the user is using RR (I did the same search in each ILS and entered “Seoul” into the search bar and 100% of the retrievals in each ILS was about the Korean capital).

The Library of Congress guide to *Searching/Displaying Non-Roman Characters* (2017) points out that “most marks of punctuation in your search query are converted to spaces. Some punctuation and diacritic marks are removed.” They also go on to explain this in their *Search/Browse Help-Searching in Korean* (2017). Since MR relies on apostrophes, this will become problematic when a space is being added instead. This causes incorrect divisions and spaces in words. One last problem related to MR’s use of diacritics is that LC warns those who submit guidelines for Romanization Tables (2010) to avoid modifier characters whose positioning over or under a Latin script base letter may interfere with printing and/or display of that letter.

Another problem with the current MR ALA-LC Romanization Table that has been pointed out by both Jeong (1998) and Kim (2006) is the problem of word division. As Kim (2006, 63-64) points out: “According to ALA-LC Romanization Tables, space should be used to separate each word or lexical unit (including noun particles) from all other words or lexical units. But this rule is contrary to Korean conventional practice which always joins particles to the nouns they modify, and often omits spaces between nouns. For instance, the title “노인 과 바다” (“The old man and the sea”) should be written as “Noin kwa pada” (“noin” = “the old man”; “kwa” = “and”) according to the ALA-LC rule. Yet Koreans usually write the title as “Noinkwa pada,” with no space between Noin and kwa. The principle of word division may create a heavy burden for Korean users in their searching processes as it is not natural. The Library of Congress and other libraries have created many faulty bibliographical records, due to incoherent and arbitrary word division.” This problem was pointed out in 2006 and when the ALA-LC Korean Romanization was updated in 2009 not only did they not feel the need to correct this confusion, they even used “Noin kwa pada” as an example under their guidelines for Section 6: Particles (조사), in which they reiterate that one must “separate a particle from other parts of speech” (2009, p. 29). Problems like this explain why Koreans users, according to a study done by Jeong (1998), were the most dissatisfied amongst CJK countries in terms of Roman Title Phrase searches. Ultimately both Koreans and non-Koreans struggle with the diacritics and word divisions of the current ALA-LC Korean Romanization Table and as a result, information is being lost or mislabeled.

**Revised Romanization Method is more effective than McCune-Reischauer**

The strengths and weaknesses of the RR Method stand in stark contrast with MR. The MR is based on phonetics, thus is great for pronouncing Korean words, the RR is not. However, the MR is full of diacritics and the RR is not, thus it is easier for information
retrieval. RR was devised in 2001 mainly because the general public never fully accepted MR. As noted above, MR liberally uses two diacritic marks: the apostrophe and the breve which made spelling inconvenient, not least because the breve cannot be reproduced on most computers. A common compromise was to ignore the diacritic marks -- a practice that often defeats the purpose of the MR system. Before the advent of RR the truth was that many people simply followed a system that existed only inside their heads (Suh, 2000). As a result the Ministry of Culture proposed a new Romanization scheme that would alleviate these problems. The new RR system was devised as a compromise between the needs of Koreans and foreigners and because it was devised with both needs in question, it is a better universal system (National Academy of the Korean Language, 2001).

Although the intentions of the new system were good, it caught the Korean public off guard, mainly because the public opinion was not taken into account. Many articles from around that time were very negative toward the new system, as was Sangwon Suh who wrote for Asiaweek (2000). In his article entitled “You Say Bulgogee, I Say Pulgogi” he wrote: “The country’s two leading English-language dailies, The Korea Herald and The Korea Times, are already refusing to adopt the new format. Whatever difficulties the authorities may have had in coming up with a new system, it looks like getting the public to accept it will be the biggest challenge of all.” However, it appears that the “authorities” have done a good job getting the public to accept the new system. Both the Korean Herald along with the Korea Times and the other Korean English-language dailies now publish their papers using the new system.

Although the RR system has become more steadily accepted and it is more convenient to retrieve information, it still has its critics. One of the main concerns is that RR is not as accurate as MR in terms of pronunciation. One such scholar who is not in favor of RR due to its pronunciation limitations is leading Korean Studies scholar Dr. James Grayson. In his essay entitled: The new government Romanization system: why was it necessary? (2006), he argues that RR is not better, if not weaker than MR. He cites examples of student’s poor Korean pronunciation when reading RR. He points to examples such as “eo” and “eu” vowel combinations used in RR which are not natural to European or American English speakers and thus confusing when trying to pronounce Korean language.

Although Romanization is important when trying to master a new language, it should not play a big role in which Romanization Table the Library of Congress chooses. Pronunciation does not play a pivotal role in information retrieval and storage. Perhaps a non-Korean person may need to read Korean script aloud in a library setting but they are not required to have perfect pronunciation. The RR system is more than adequate for pronunciation. In fact, even the MR system is not a perfect pronunciation Romanization system. “The MR system is an inconsistent scheme for bibliographic records because Korean has several local dialects, and there are unique pronunciations in each local dialect. In other words, phonetic-based Romanization in bibliographic records might be different according to different catalogers” (Kim, 2006).
In regards to Romanization and pronunciation in general the Library of Congress now admits that "Any future ALA/LC Romanization Tables should be transliteration schemes rather than schemes to replicate pronunciation or guides to pronunciation. Pronunciation is variable around the world" (Library of Congress, 2010). As a result, pronunciation should not play any type or role when selecting an official Romanization scheme for a library.

Other people have argued that RR is not fully consistent yet and the system is still kind of new. In response to these opinions, the Ministry of Culture and the Korean Government are doing what they can to standardize Romanization both in Korea and abroad. One such way is the standardization of Korean food. The Ministry of Culture worked closely with the Ministry of Agriculture, Food and Rural Affairs to produce a draft that includes the Romanization as well as suitable translations of about 200 different Korean dishes into English. Both ministries are pushing for Korean restaurants in Korea and abroad to use the standardized terms proposed in the draft which follows the RR method (with the exception of kimchi which is already too well and entrenched in dictionaries to change, under RR it should be spelt gimchi), instead of using their own methods of Romanization and translation. What is notable about the draft is that alongside bulgogi, kimchi and bibimbap, dishes that appear in the dictionary in their Romanized form, the Romanized “gimbap” as opposed to “kimpab” (MR), will be included in the Oxford dictionary (Cha, 2013).

Another reason that people are slow to fully endorse the RR system has to do with personal names. This has always been a problem and still is, as Jaeyeong Woo (2014) shows in his article Koreans await Kim Yu-na, or is it Yuna Kim? published in the Wall Street Journal. The current official system of Romanization adopted by the Ministry of Culture in 2000 states that "surnames are not required to follow the new system." This has created a situation in which long used popular forms, the current RR system, and the MR system are all being used for surnames based on individual preference (S. Lee, 2013). As a result Korean people usually Romanize their surnames following the most popular Romanized spelling, and their given names according to personal preference. For example, ninety-five percent of all people with the surname “이” write their surnames as “Lee.” Yet according to the RR system, “이” should be Romanized as “I” and “Yi” in MR (Kim, 2006). The RR system is more open by allowing Koreans to choose for themselves and as a result one specific variation can be established. MR on the other hand claims a person’s given name should be transcribed according to the system’s rules, and not in the way that he or she prefers to spell his or her name. As Sun Kyung Kim in her article “Romanization in Cataloging of Korean Materials” (2006) points out: “despite the rigidity of these rules, looking through the RLIN and the Library of Congress authority files and bibliographic records, it can be seen that even catalogers make mistakes when Romanizing Korean names. Many records in RLIN and the Library of Congress have inaccurately spelled names.” Kim goes on to use Lee, Duhyun, a famous Korean author, as an example. Lee, Duhyun, should be Romanized as “Yi, Tu-hyon,” according to the MR system. In the UCLA catalog, however, the user can find three items under “Yi, Tu-hyon,” and fourteen entries can be found under “Lee, Duhyun.” The authorized
form of this name in LC/NACO Name Authority File (NAF) is also “Lee, Duhyun, 1924-.” These results show that the MR system is inconsistently applied when catalogers create bibliographic records.

The Library of Congress supports MR, which has a stringent standard that confuses catalogers. Because RR is not as strict, it allows catalogers to follow cataloging rule AACR2 22.3C2 (Names written in a non-Roman script) which surmises that you should “choose the Romanized form of name that has become well-established in English language reference sources for a person entered under surname whose name is in a language written in a non-Roman script” (ALA, 2002). As a result, a cataloger can easily use VIAF (Virtual International Authority File) or *Academic American Encyclopedia* to establish the commonly recognized form of the name, thus further standardizing CJK cataloging.

In 2003 the Korean Government sanctioned Seoul National University to gather opinions and conduct research on the issue with the aim of creating a reasonable system for Romanizing Korean surnames. The findings, as reported by Sang-Oak (Timothy) Lee, showed that approximately 72% of the participants of the survey indicated opposition to the use of diacritics. In respects to what system Koreans favored, the study showed that regarding the degrees of preference for the four different forms of Romanization (RR, MR, MY [Martin Yale], grafts of RR and MR together) the 2001 proposal of the National Institute of the Korean Language (RR) received the highest degree of support, more than for any of the other proposals. The study also showed RR has the highest degree of recognition amongst the four Romanization forms.

Lastly, leaving the spelling of Korean names up to themselves is not such a far-fetched idea. In America there is no standardized way of spelling a name. In fact, Fetzer (2013) shows that amongst the Social Security Administration’s top 1,000 baby names there are seven different spellings for Cameron (Camren, Camron, Camryn, Kameron, Kamron and Kamryn) and 9 different spellings for Kayln (Cailyn, Kaelyn, Kaelynn, Kailyn, Kailynn, Kaylen, Kaylyn, Kaylynn). These are just two of many examples. Americans also choose various spellings for personal names. In the end, the RR system of Romanization is more flexible than MR and follows more in line with the rest of the world.

**A Similar Transition has Proven to be Successful in the Past**

The Library of Congress shifting from a pronunciation based Romanization scheme to a more user-friendly scheme is not unheard of, it has happened in the past, most recently in 1997, in regards to Chinese Romanization. As Arsenault (2002) explained that prior to 1997 two romanization systems for Chinese data were currently in use in most libraries in the Western World: the Wade-Giles (WG) system, mainly used in North American libraries, and pinyin--called Hanyu pinyin but simply referred to as pinyin--mainly used in European and Australian libraries. Which is similar to the case of Korean Romanization whereas America relies on MR whereas the rest of the Western World uses RR. There was a consortial effort between LC, Research Library Groups (RLG) and OCLC to determine if there was not
only a legitimate need to switch Romanization schemes but whether this was economically feasible. A lot of research was conducted to find the best solution.

Some history on WG and pinyin: before the late 20th century, the most widespread Romanization scheme in China was the WG, which was adopted by the Library of Congress to Romanize Chinese-language materials. WG was created in 1859 by Sir Francis Wade (1845-1935) and revised by Herbert Allen Giles (1845-1935), his successor in 1892. WG was the first standard for the Romanization of Chinese (Perkins, 2013). However, in 1958 the government of the People's Republic of China (PRC) approved pinyin as the new Romanization system for Mandarin, yet the Library of Congress continued to use WG. “The Library of Congress, in 1979 and 1980, recommended that the library community undertake conversion from Wade-Giles to pinyin for the Romanization of Chinese. It was anticipated that "more and more people will, in the future, approach Chinese through pinyin Romanization..."; "... fewer and fewer library users will have a working knowledge of Wade-Giles." The library community, however, voted to retain Wade-Giles” (Lu, 1995). In this case, the Library of Congress recognized the need for transition but libraries opposed this transition because it would be difficult, time consuming and expensive. Research was continuously undertaken to determine the feasibility of a switch. By 1997, pinyin was already the standard of the West including the US government for two decades and the United Nations. Libraries were virtually the only institute in America still using WG (Melzer, 1997). The Library of Congress and the libraries stood apart from the rest of the world.

In 1997 the Library of Congress announced that “in order to provide better service to library users, the Library of Congress is moving to adopt the pinyin system of Romanization of Chinese” (Library of Congress, 1998). One of the main reasons for this was because already at that time most users of American libraries were more familiar with the pinyin Romanization of Chinese names and places, providing access to the Chinese language with that system would make it easier for them to locate material. The use of pinyin Romanization by libraries was expected to facilitate the exchange of data with foreign libraries (Library of Congress, 2007). The same can be said about Korea. Most users of American libraries are more familiar with the RR of Korean names and places and by switching to RR, it should be easier to facilitate the exchange of data with foreign libraries.

The decision to adopt pinyin over WG did not happen overnight, even when it became apparent that pinyin was more convenient. Neither did the transition from WG to pinyin as the official the Library of Congress Romanization happen overnight. They followed these steps:
1. LC conversion of Chinese authorities by OCLC
2. Conversion of LC bibliographic records (bibliographic records) by Research Library Groups (RLG)
3. Bibliographic records conversion by OCLC and RLG of their respective union catalogs
4. Conversion by OCLC of the non-Chinese records containing Chinese text, and later by RLG of similar records in their databases
5. Conversion efforts by OCLC and RLG of records of institutions from WG to Pinyin (Thornburg, 2002). It was a process but one that has proven to be successful and can happen in regards to Korean Romanization too.

Another reason many people wanted to switch from WG to pinyin mirrors the reason that the Library of Congress should consider switching from MR to RR; namely a change would make information retrieval easier. Already in 1980, scholars were requesting a change. “Pinyin is a much more practical system than Wade-Giles, the most commonly-used transcription system in the West. It was designed with minimum recourse to the use of diacritics, eliminating the use of hyphens between syllables and apostrophes after certain consonants as well as the raised tone numbers of the Wade-Giles system...These features make it not only eminently suitable for printing and typewriting purposes, but also very easy to write cursively and to read” (Chappel, 1980, 112). Just like MR, WG uses diacritics which cause confusion and errors so pinyin was created and began to curry favor amongst information professionals. And as was pointed out earlier, omitting diacritics causes great errors in retrieval in regards to information retrieval when using MR. The same can be said about using WG. As Arsenault (1998) pointed out; previous analysis revealed that, in indexing, due to systematic removal of diacritics and punctuation, two hundred of the 410 WG base syllables (or toneless syllables) are collapsed into one hundred syllables. For example, pa and p’a are both searched with the string "pa." This represents a loss of one hundred syllables. Furthermore, twelve base syllables are collapsed into three syllables-chu, ch’u, chu, and ch’u are all indexed as "chu"; the same for chun and chuan--which represents an additional loss of nine base syllables, for a total of 109 syllables lost. The remaining 198 syllables remain unaffected. These indexing problems are also prevalent went using MR.

Results of a survey of East Asian Studies institutions done in 1997 by Amy Tsiang, then Head Librarian of UCLA’s Richard C. Rudolph East Asian Library, showed that 84% agreed that Wade-Giles Romanization should be converted to pinyin. It was research like this that finally convinced the Library of Congress to finally adopt pinyin as the official Romanization scheme of the Library of Congress. The Library of Congress commenced the Pinyin Conversion Project in which they would slowly (1997-2001) implement the changes. Already, research from 2002 (Arsenault) showed that the transition to pinyin was working and was necessary. The research showed that based on high success rates measured for retrieval tasks “indicate that Romanization is, in most cases, a fairly effective means for retrieval of Chinese language titles” the research further indicated that the use of wrong Romanization only occurred in the group using Wade-Giles. In addition to this, the adoption of the pinyin system contributed to improving bibliographic access (Kim, 2006, 71).

One last example to show how even scholars can become confused when two separate Romanization schemes are being promoted; Dr. George J. Leonard, Professor of Interdisciplinary Humanities at the University of San Francisco, shared a story where a
respected English professor once confidently remarked to an audience that he had been reading Qing Dynasty poetry. He pronounced it, Kwing. Presumably he thought the Qing Dynasty and the Ch’ing Dynasty were two different dynasties. They are not; they are both, simply, Ch’ing. He also points out that the “Zhou” and “Chou” is the same dynasty as well (Leonard 2011). Just like in Korea, having a state sponsored Romanization scheme that differs from the Library of Congress causes a lot of confusion for many people, scholars included.

Although RR is not used as universally as pinyin was at the time of transition, and the RR does not have the longevity of pinyin, there are similarities between RR and pinyin. MR follows the old Wade-Giles systems of Chinese in that the criterion for the phonetic values of the letters of the Latin alphabet is the vowels as in Italian and the consonants as in English (Royal Asiatic Society 1961). As a result, just like with pinyin, RR was needed to simplify Romanization. Just like in China, Korea learned that diacritics in Romanization are not effective. Just like in China, the Library of Congress should consider a transition from a phonetically based Romanization scheme full of diacritics to one that is more user-friendly, convenient and easier to use in regards to information retrieval. At the time when LC adopted WG over pinyin, most users were familiar with pinyin due to it being the international standard and it was also used in publications such as dictionaries and maps, and sometimes for book and periodical titles. It was widely seen in public places such as building names; street, highway, and railway signboards; and on product labels (DeFrancis, 1990). The same can easily be said about Korea. Most people see RR being used in dictionaries, maps, streets, product labels, etc. Romanization remains a fairly effective means for retrieving Asian language titles, but it only effective if the system is free from diacritics and is one that is recognized by the majority of the population.

Conclusion

The research for this paper shows how difficult search retrieval using Romanized Korean words can be due to variant official Romanization schemes (the Korean government officially recognizes their own Revised Romanization method while the Library of Congress supports McCune-Reischauer). Somebody retrieving information using Romanized Korean words would need to be familiar with both Romanization schemes and would need to know how to use Advanced Search options in order to retrieve precise results. The question then is how can this be fixed? The ALCTS Non-English Access Working Group on Romanization Report (2009) shows why Romanization is still necessary in libraries so eliminating Romanization completely is not a viable option yet. One possible solution is for the Korean Government and the Library of Congress to agree on one standard Romanization scheme. Jeong (1998) and Kim (2006) along with other studies have shown that the use of diacritics in MR causes confusion and irritation. As a result, RR is a more convenient method for information retrieval. China experienced a similar situation in regard to Chinese Romanization. Their old system of Romanization, Wade-Giles, also used diacritics and was
replaced by pinyin, which was more convenient for users retrieving information using Romanized Chinese words. But this process took a long time and was costly. So ultimately the lingering question is whether the Library of Congress needs to formally recognize RR as the official the Library of Congress method of Korean Romanization. As of right now, there has not been enough research done to warrant a decision either way. Just like the Library of Congress did in regard to the conversion from WG to pinyin, they should support active research into whether a transition from McCune-Reischauer to the Korean Revised Romanization method is necessary.

References


74


75

Appendix A: Joseon Dynasty and Chosŏn Dynasty in Google

Joseon Dynasty (Google)

2. Joseon Dynasty – Ancient Korean History – Asian History asianhistory.about.com
4. Royal Tombs of the Joseon Dynasty – UNESCO World ... whc.unesco.org
5. Joseon Dynasty – New World Encyclopedia www.newworldencyclopedia.org/entry/Joseon_Dynasty
7. Korean History 8/8 Joseon Dynasty 1392~1910 (조선 왕조 ... www.youtube.com/watch?v=bPAj7ft4jU0
8. Joseon Dynasty Fashion – YouTube www.youtube.com/watch?v=9t9jszwwcb0
9. The Last Dynasty of Korea 조선 – YouTube www.youtube.com/watch?v=nVwVEQFnoes
10. Treasures from Korea: Arts and Culture of the Joseon ... www.mfah.org/.../treasures-korea-arts-and
11. Treasures from Korea: Arts and Culture of the Joseon ... www.philamuseum.org/exhibitions/795.html
12. In Pursuit of White: Porcelain in the Joseon Dynasty, 1392 ... www.metmuseum.org/toah/hd/.../hd_chpo.h
13. Treasures from Korea: Arts and Culture of the Joseon Dynasty www.lacma.org/art/.../last-dynasty
14. Rulers of the Joseon Dynasty and Kdrama Interpretations ... https://thetalkingcupboard.wordpress.com/.../rulers-of-the-joseon-dynast
15. Treasures from Korea: Arts and Culture of the Joseon ... www.amazon.com/.../Religious/.../Buddhism
16. The 40 Royal Tombs of the Joseon Dynasty – Korea Tourism ... asiaenglish.visitkorea.or.kr/ena/CU/CU_EN_8_4_2_8.jsp
17. Annals of the Choson Dynasty – 조선왕조실록의 홈페이지에 ... sillonk.history.go.kr/etc/ashing.jsp
Chosŏn Dynasty (Google)

1. Joseon – Wikipedia, the free encyclopedia (1)
en.wikipedia.org/wiki/Joseon
2. Choson dynasty | Korean history | Encyclopedia Britannica (3)
www.britannica.com/topic/Choson-dynast
3. 1450-1750: Korea: The Choson (Yi) Dynasty | Central ...
afe.easia.columbia.edu/main_pop/kp_choson.htm
4. Joseon Dynasty – Ancient Korean History – Asian History (2)
asianhistory.about.com/South Korea (Republic of Korea)
5. Choson Dynasty – Minneapolis Institute of Arts
archive.artsmia.org/art-of-asia/history/korea-choson-dynasty.cfm
6. Joseon Dynasty – New World Encyclopedia
www.newworldencyclopedia.org/entry/Joseon_Dynasty
7. Royal Tombs of the Joseon Dynasty – UNESCO World ... (4)
whc.unesco.org
8. Korean History – Early Choson Period – AsianInfo.org
www.asianinfo.org/asianinfo/korea/history/early_choson_period.htm
9. Korean History 8/8 Joseon Dynasty 1392~1910 조선 왕조 ...
www.youtube.com/watch?v=bPAj7ft4jU0
10. Korea’s Joseon Dynasty, to 1700 – Macrohistory
www.fsmitha.com/h3/h24kor.html
11. In Pursuit of White: Porcelain in the Joseon Dynasty, 1392 ...
www.metmuseum.org/toah/hd/Chpo.h
12. KoreanHistory.info Choson or Joseon Dynasty ?? 1392 – 1910
koreanhistory.info/ChosonDynasty.htm
13. Chosŏn | Board Game | BoardGameGeek
boardgamegeek.com/boardgame/161943/chosn
14. The Annals of King T’aegio: Founder of Korea’s Choson ...
www.amazon.com/The-Annals-King.../0674281306
15. The Origins of the Choson Dynasty – Amazon.com
www.amazon.com/Origins-Dynasty
16. South Korea – The Choson Dynasty – Country Studies
countrystudies.us/south-korea/5.htm
17. Yi Ku, the Last Prince of the Choson Dynasty » Center for ...
www.bu.edu/yi-ku-the-last-prince-of-the-choson-dy
18. The Origins of the Choson Dynasty – University of Washington
www.washington.edu/uwpress/DUNORI.htm
19. An Introduction to the Joseon Dynasty (1392–1910) | Asian ...
education.asianart.org/explore/introduction-joseon-dynasty-1392-1910
20. The Neo-Confucian Foundation of The Chosŏn Kingdom ...
Appendix B Joseon Dynasty and Chosŏn Dynasty in WorldCat

Joseon Dynasty (WorldCat)

1. **Treasures from Korea : arts and culture of the Joseon Dynasty, 1392-1910 (2014)**
   by Hyunsoo Woo
2. **In grand style : celebrations in Korean art during the Joseon dynasty** (2013)
   by Hyonjeong Kim Han
3. **Pathways to Korean culture : paintings of the Joseon Dynasty, 1392-1910** (2014)
   by Burglind Jungmann
4. **Celebrating events with banquets and ceremonies in the Joseon dynasty** (2011)
   by Chŏng-hye Pak
5. **Visits to the kings : guidebook of royal tombs of the Joseon Dynasty** (2008)
   by Byeongyu Lee
   by Jackie Menzies
   by Kungnip Munhwajae Yŏn’guso
8. **Paintings of late Joseon Dynasty** (2006)
   by Leuem
9. **The beautiful memories of the Joseon dynasty : National Palace Museum of Korea, the 50 highlights.** (2010)
   by Kungnip Kogung Pangmulgwan
    by Hoam Misulgwan
    by Han’guk-Kukche-Kyoryu-Chaedan
    by Asian Art Museum of San Francisco
13. **Course Deoksugung : Where the Joseon Dynasty ended and the history of the Korean Empire began.** (2014)
    Jo MyeongHwa
    by Kumja Paik Kim
15. **조선 시대 진연 진찬 진하 병풍 / Folding screens of court banquets and congratulatory ceremonies in the Joseon dynasty / by So In-hwa, Park Jeong-hye, Judy Van Zile.**
    Chosŏn sidae chinyŏn chinch’an chinha pyŏngp’ung / Folding screens of court banquets and congratulatory ceremonies in the Joseon dynasty / by So In-hwa, Park Jeong-hye, Judy Van Zile. (2000)
   by Francis D K Ching
17. *Interior space and furniture of Joseon upper-class houses* (2011) 
   by Sang-hoon Ch’oe
   by Seunghee Kim
19. *白磁を飾る青: 朝鮮時代の青花* = Blue-and-white ware of the Joseon Dynasty 
   = 백자를 장식한 青: 조선시대의 靑畵 / 
   *Hakuji o kazaru ao: chōsen jidai no seika* = Blue-and-white ware of the Joseon Dynasty 
   = Paekja rul Changsik han chong : Choson sidae ui chonghua (2010) 
   by Ōsaka Shiritsu Tōyō Tōji Bijutsukan
   by Yunesŏk’o Han’guk Wiwŏnhoe

Chosŏn Dynasty (WorldCat)

   by Peter H Lee
   by John B Duncan
   by James B Palais
   by Mary E Connor
   by Byonghyon Choi
   by Hyunsoo Woo
   by Michael Brand
8. *The Kinyo institution in Chŏson dynasty Korea, as seen in the Chosŏn-wangjo-sillok: (veritable records of the Chŏson dynasty)* (2001) 
   by Vincenza D’Urso
9. *In grand style: celebrations in Korean art during the Joseon dynasty* (2013) 
   by Hyonjeong Kim Han
10. *Rapt in colour: Korean textiles and costumes of the Chosŏn dynasty* 
    = 조선시대의한국의상과보자기 / 
    *Rapt in colour: Korean textiles and costumes of the Chosŏn dynasty* = Chosŏn sidae ŭi Han’guk ŭisang kwa pojagi (1998) 
    by Claire Roberts
    by Burglind Jungmann
by Po-gi Son
by Gorō Akaboshi
by Kumja Paik Kim
15. South Korea: a country study (1992) 
by Andrea Matles Savada
by Hoam Misulgwan
by Michael Wickman
18. Click into the Hermit Kingdom: virtual adventure into the Choson dynasty (2000) 
by Sung-jin Yang
19. Adoption and transformation of the Mi tradition in landscape painting of the late Choson dynasty (1700-1910) (1993) 
by Hyung-min Chung
20. Decorative designs in the houses of Chosun Dynasty period (1988) 
by Pyŏng-ok Ch’ŏn

Appendix C Seoul and Sŏul in Google

Seoul (Google)

1. Home | The Official Website of Seoul Metropolitan Government english.seoul.go.kr/
2. Seoul - Wikipedia, the free encyclopedia en.wikipedia.org/wiki/Seoul
7. Seoul Tourism: Best of Seoul, South Korea - TripAdvisor www.tripadvisor.com › Asia › South Korea
8. Seoul Time - Local Time in Seoul - Time Zone in Seoul localtimes.info › Asia › Korea (south)
9. Seoul’s stream on SoundCloud - Hear the world’s sounds https://soundcloud.com/seoulmtl
10. Seoul | Official Korea Tourism Organization english.visitkorea.or.kr/mapInfo.kto?func_name=depth2...1
11. Seoul – Travel guide at Wikivoyage
   en.wikivoyage.org/wiki/Seoul
12. SEOUL SEMICONDUCTOR
   www.seoulsemicon.com/
13. Seoul Weather - AccuWeather Forecast for Seoul South Korea
   www.accuweather.com › ... › Asia › South Korea › Seoul
14. Del Seoul Korean Street BBQ
   delseoul.com/
15. Downtown Seoul hotels – brand new Courtyard ... - Marriott
   www.marriott.com/.../selcy-courtyard-seoul-times-s...
16. Seoul | CNN Travel
   travel.cnn.com/seoul
17. Seoul National University
   www.useoul.edu/
18. Seoul: 10 Things to Do — Introduction - TIME
   content.time.com/.../0,31489,1848378_1848364_1847892,00.html
19. Seoul (@SeoulMTL) | Twitter
   https://twitter.com/seoulmtl
20. Seoul, South Korea Weather Forecast and Conditions ...
   www.weather.com/.../Seoul+South+Korea+KSXX0

Sŏul (Google)

1. Seoul - Wikipedia, the free encyclopedia
   en.wikipedia.org/wiki/Seoul
2. Sŏul - Wiktionary
   en.wiktionary.org/wiki/Sŏul
3. Sŏul, South Korea - Sunrise, sunset, dawn and dusk times ...
4. Concerts at 잠실실내체육관, Sŏul – Last.fm
   www.last.fm/venue/8805169+잠실실내체육관
5. Sŏul Special City, South Korea, Drama Korea Lie to Me и ...
   https://www.facebook.com/pages/Sŏul...to.../145630412286448
6. Sŏul-t’ŭkpyŏlsi, South Korea - Geographic.org
   www.geographic.org/geographic_names/name.php?uni...fid...c
7. Sŏul-t’ŭkpyŏlsi on Cowbird
   cowbird.com/place/kr:so/
8. sŏul-t’ŭkpyŏlsi | Tumblr
   https://www.tumblr.com/tagged/Sŏul-t’ŭkpyŏlsi
9. Plan a trip to Sŏul-t’ŭkpyŏlsi, Asia - TravelMuse
   www.travelmuse.com/destinations/KR/11/00-soul-t-ukpyolsi
10. Lexington Hotel in Seoul – 4 star hotel in Seoul - HRS
    en.hrs.com › South Korea ›
11. The Journal of Korean Studies, Volume 19, Number 1 (Spring ...
13. Sŏul üi sanŏp kujo pyŏnhwa wa... - HathiTrust
catalog.hathitrust.org/Record/006247878
14. Reassessing the Park Chung Hee Era, 1961-1979: ...
https://books.google.com/books?isbn=0295801794
15. Che 13-hoe Sŏul Yŏn’gŭkche The 13th Seoul Theatre ...
www.loc.gov › ... › Prints & Photographs Online Catalog
16. Romeo and Juliet Yong-ch’ŏl Pak Sŏul-si Korean Edition ...
www.amazon.com/Juliet-Yong-cho.../8973812084
17. Publisher: parhaengch’ŏ Sŏul T’ukpyŏlsi (Open Library)
https://openlibrary.org/.../parhaengch’ŏ_Sŏul_T’ukpyŏlsi
18. Chu bu dao de ai qing Hollidei in Sŏul = Holiday in Seoul,
libcat.parracity.nsw.gov.au/.../ada
19. Hotel RA HTL SUITES SEOUL NAMDAEMUN in Seoul – 5 ...
hrs.de › Südkorea › Seoul
20. tosi munhwa sidae ŭi Sŏul ŭl kiri’nun ch’aek = Seoulness
catalogue.nla.gov.au/Record/1446687

Seoul (WorldCat)

   by Martin Robinson
2. Seoul (2001)
   by Maps.com
   by Chae-wŏn Kim
   by Robert Koehler
5. Seoul (1999)
   by Robert Storey
   by Martin Zatko
7. Seoul.
   Journal, magazine. Lonely Planet Publications ©1996-
   Time Out Guides, 2011.
   by Robert Storey; Chris Taylor
10. Seoul.
    by Chung-ang Atlas Co.
by Rose E Lee; George Mitchell
    Print book GeoCenter [distributor], 2010.
    by Falk-Verlag (map)
    by DK Publishing, Inc.
    by TravelVideoStore.com
    by Peter Hyun
    Journal Seoul Selection
    by Periplus Editions (map)
    Journal, magazine Avalon Travel, ©2009-Database: WorldCat

Sŏul (WorldCat)

1. The rise and progress of religion in the soul Illustrated in a course of serious and practical addresses ... with a devout meditation or prayer added to each chapter (1756)
   by Philip Doddridge
2. The life of God in the soul of man : or, the nature and excellency of the Christian religion. With The Methods of attaining the Happiness it proposes: Also an Account of the Beginnings and Advances of a Spiritual Life. With a preface. By Gilbert Burnet, now Lord Bishop of Sarum. (1702)
   by Henry Scougal
   by Seal
   by Inazo Nitobe
5. Chicken soup for the soul : 101 stories to open the heart & rekindle the spirit (1993)
   by Jack Canfield
6. The soul (200)
   by William, of Auvergne Bishop of Paris
7. Soul. (1970)
   by Lee Rainwater
8. Soul on ice. (1968)
   by Eldridge Cleaver
   by Tobsha Learner
10. Soul.
    Journal, magazine Soul Publications
by Andreï Platonovich Platonov
12. The soul (1994)
   by Adrian Kuzminski
    by Bob Brunning
    by Lena Horne
15. Soul (1992)
    by British Broadcasting Corporation
    by Kentucky HeadHunters
17. Soul (1993)
    by Margaret Becker
    by Erikka Haa
    VHS video Century Home Video
    by H. G. Wells
Building a “One-hour Library Circle” in China’s Pearl River Delta Region

Paul W.T. Poon
University Librarian
University of Macau

In early June 2017, the University of Macau, Sun Yat-sen University in Guangzhou, and the Chinese University of Hong Kong signed an agreement on the establishment of the Guangdong-Hong Kong-Macau University Library Alliance (粵港澳高校圖書館聯盟), officially ushering in the birth of a “One-hour Library Circle” in Pearl River Delta Region in Southern China. Dr. Paul W.T. Poon, former University Librarian of City University of Hong Kong, and currently University of Macau Librarian since 2005, was elected the Founding President of the Alliance. The University Librarians in Sun Yat-sen University and the Chinese University of Hong Kong were named Vice Presidents. The libraries of these three universities are, therefore, the founding members of the Alliance, and they will then invite other university libraries in their region to join as members.

Inspired by the win-win situation created by the establishment of the Guangdong-Hong Kong-Macau University Alliance (粵港澳高校聯盟) in late 2016, with 26 universities from the three regions as participating members, the University of Macau Library initiated the process of forming a library alliance under its parent alliance. This proposal was agreed with by SYS and CUHK libraries. The mission of the Library Alliance is to engage the major university libraries in the Pearl River Delta region and optimize the use of academic resources through deep collaboration, thus rendering support to teaching, learning, and research – the three important tasks being carried out in the universities.

Member libraries will work together to implement resource building and sharing, promote scholarly communication, and make the most of existing and emerging technologies to enhance library efficiency and further the cooperation among the major universities in Guangdong, Hong Kong and Macau. The primary tasks of the Library Alliance will be focused on interlibrary loan, document delivery, exchange of publications, exchange and training of library personnel and holding professional seminars and conferences.

Looking forward to future activities, the University of Macau Library, on behalf of the Library Alliance, will organize and host an international symposium on the best practice of library consortium along with the first Library Directors’ Forum at the end of 2017. Experts from countries such as Portugal, the United States and Australia will be invited as guest speakers to share their experiences and insights on library regional collaborations and opportunities. This activity will be fully supported by the Tertiary Education Services Office of the Macao SAR Government. Other future tasks that the Alliance will undertake include producing its website and tackling other housekeeping chores. The long-term goal of the Library Alliance is to reach out to other similar alliances in the nation, such as the Yangtze
River Delta University Library Alliance (長三角高校圖書館聯盟), the Pearl River Delta Digital Library Alliance (珠江三角洲數字圖書館聯盟) and the Alliance of Provincial Public Libraries in Western Areas (西部省級公共圖書館聯盟) so as to create a win-win scenario for all parties concerned.

Finally, as the Guangdong, Hong Kong and Macau municipal governments are now developing a Greater Bay Area (大灣區) to foster regional integration in southern China, it is hoped that the Library Alliance shall seize this opportunity to make contribution to the success of the Greater Bay Area by providing support to high-level research work and creative innovation.

For more details of the Library Alliance, please contact Dr. Poon at wtpoon@umac.mo.

Photo caption: The three founding libraries signed an agreement and celebrated the establishment of the Library Alliance on 6 June 2017.
Interview with the Curator of the Po Leung Kuk Museum

Dr. Patrick Lo
Associate Professor, Faculty of Library, Information & Media Science, University of Tsukuba

Dr. Dickson Chiu
Lecturer, Faculty of Education, University of Hong Kong

Introduction

Founded in 1878, the Po Leung Kuk (PLK) aimed at protecting basic welfare and social justice of children and women in Hong Kong. Having witnessed the history of Hong Kong's economic and sociocultural growths in the past centuries, PLK's functions and services have now been extended to education, child care, elderly care, and many other aspects of public services. The Po Leung Kuk Museum currently maintains a large number of archives and artifacts with high historical and cultural values. In the following interview, Rosana Leung, Curator of Po Leung Kuk Museum explains the historical background, importance and uniqueness of the PLK museum archives collection, as well as sharing with the readers her unique management experience in archival science in the museum context.

Could you please give a brief account of the history of the Po Leung Kuk (PLK)?

In mid-19th century, kidnapping and human trafficking of women and children was a rampant crime in Hong Kong. Women were frequently sold as domestic servants or prostitutes, and children were sold as adopted sons and girls. These victims would then be shipped to Southeast Asia or farther location. As the trend of human trafficking grew unstoppable, several Chinese merchants resided in Hong Kong from the county of Dongguan (東莞) petitioned to the Governor of Hong Kong, Sir John Pope Hennessey, requesting the adoption of anti-kidnapping policy in November 1878. The motive behind these Dongguan merchants was that many of the victims of kidnapping cases were from their hometown, thus the merchants were eager to protect the welfare of their own
people. After negotiations with the Government and the Colonial Office, Po Leung Kuk\(^1\) was established in Hong Kong as a charitable organisation dedicated to anti-kidnapping work and provision of shelter to victims of kidnappers.

Throughout the 20\(^{th}\) century, the PLK’s mission evolved from assisting the kidnapped into a full-fledged charity which now provide a wide spectrum of services, from social, educational, recreational, cultural to medical.

**When was the PLK Museum Archives established?**

The history of Po Leung Kuk Museum can be traced back to 1979. Encouraged by Dr. Elizabeth SIN Yuk Yee, the Kuk began to organise an archive for systematic storage and compilation of documents and information. In 1984, the Po Leung Kuk Archives was formally established and received scholars from the US and China. In 1998, PLK Archives was expanded and opened to the public. In 2002, the Archives was once again expanded and was renamed as Po Leung Kuk Museum. Since then, it has been providing thematic exhibitions to promote public understanding in our history and services.

**What were the original aims and purposes for establishing the PLK Archives?**

The Po Leung Kuk Museum is responsible for the acquisition, conservation and research of historical documents and artifacts related to Po Leung Kuk. As a museum, it also provides general and thematic exhibitions of the Kuk’s history in order to foster a better appreciation of the Kuk’s development and relationship with the history of Hong Kong.

**What is the current size, format, contents and nature of the PLK Archives?**

The Po Leung Kuk Museum houses a great variety of documents and collections, ranging from records, ledgers, statements and director meeting minutes to plaques, awards, trophies, certificates and other historically significant items related to the Kuk’s history. Currently, the Museum aims to enrich its wealthy collection in order to provide a holistic perspective of the Kuk’s history in all historical periods.

**How is the PLK Archives collection related to the history of Hong Kong?**

Po Leung Kuk is a renowned charity in Hong Kong and enjoys a history that has spanned nearly the entire colonial period of Hong Kong. Committed to the suppression of

---

kidnappers and care of the kidnapped, Po Leung Kuk has accumulated a huge amount of data regarding migrant pattern, human trafficking, the *mui tsai* (妹仔) issue and connections between prominent Chinese leaders and the colonial government. These data may reflect the intertwined relationship between the Kuk and the Hong Kong society, as many of the Kuk’s services, such as nurseries and elderly homes and centres, were created to respond social demands of the time. Such records are precious in learning the social history of Hong Kong, and scholars may take full advantage of our well-organised, rich collections for their related research in Hong Kong history.

For the general public, the Museum frequently organises thematic exhibitions inspired and referenced from its archives to provide unique and interesting experience in learning the history of Hong Kong’s development from a small port city into a modern metropolis. Multiple workshops, and seminars were also held each year to promote appreciation in conservation of cultural heritage.

**What kind of skills and knowledge are needed for managing the PLK Archives?**

Of all knowledge and skills, we first of all valued most importantly the employee’s interest in the history of Po Leung Kuk. PLK’s history, compared to other general history in Hong Kong, is a narrow topic. In order to be motivated into learning, digging into, and researching our huge archives, the staff must possess a serious interest, and solid familiarity in the history of Hong Kong in order to contribute to our conservation and preservation effort. While our Museum is not as large as government-funded museums, the Museum remains competitive through its unique collections, and contributes a crucial element in the whole picture of Hong Kong’s history.

**What are the major challenges currently faced by the PLK Archives Manager and staff?**

Unrealsied by the majority of the public, the PLK was and still is an enormous player in shaping the social history of Hong Kong. Thus, it has been accumulating a great amount of documents and records that are invaluable in learning the past of Hong Kong. For our management, it is always a challenging task to introduce an indefinite amount of our history to the public with a definite amount of manpower and time. Such difficulty is not only a problem for small museums but also for large museums across the world, and

---

2 *mui tsai* (妹仔) “little sister” in Cantonese: a South China practice whereby young girls would be acquired by a family to work for them and later married off.
our Museum shall continue to overcome such issues through experience, and creativity in order to connect more visitors across the globe.

**Who are the majority users of the PLK Archives?**

The Museum has been open to the public since 1998, and has received more than 123,200 visitors (by August 2015). Many of the visitors came from local organisations and institutes, such as primary schools, secondary schools, colleges, universities, elderly care centers and other social units commonly found in Hong Kong. Guests such as corporate representatives, government officials and foreign dignitaries are also frequent sights in the Museum. On the other hand, the Museum has been providing information access to scholars across the world since 1984, such as those from Australia, Japan, Singapore and the United States.

**What kind of reference questions do you usually receive at the PLK Archives?**

As the development of Po Leung Kuk is closely related to the social history of Hong Kong and the history of Chinese migrant workers, scholars frequently accessed information related to the history and records of admitted women and children from the late 19th century to the interwar period. Many scholars wrote their articles and books on the lives of *mui tsai*, ordinary women and migrants with reference to Kuk records. Some researchers also chose to access our archives in order to analyse the relationship between wealthy and prominent Chinese merchants and the colonial government between late 19th century and early 20th century, mostly due to the fact that Po Leung Kuk directors enjoyed great prestige and social respect in the aforementioned period. In addition, the general public is greatly interested in the Kuk’s adoption services and the operation of homes for children, while they remain curious about the overall history of the Kuk.
Web-Archiving Chinese Social Media:
Final Project Report
August 2017

Yunshan Ye (East Asian Studies Librarian, Johns Hopkins University), Project Lead
Ding Ye (Asian Studies and Linguistics Bibliographer, Georgetown University)
Cathy Zeljak (Director of Global Resources, George Washington University)
Daniel Kerchner (Senior Software Developer and Librarian, George Washington University)
Yan He (China Studies Librarian, George Washington University)
Justin Littman (Software Developer and Librarian, George Washington University)

I. PROJECT OVERVIEW

Recognizing the critical roles of social media in Chinese socio-political activism and importance and the urgency of preserving these born digital, ephemeral records, librarians from three universities, including Johns Hopkins University (JHU), George Washington University (GW), and Georgetown University (GU) decided to join forces and collaborate on a web-archiving project in 2014. Using the latest technology available, the project intended to preserve Chinese social media responses to the ongoing anti-corruption campaign, a major political campaign launched by the Chinese President Xi Jinping in 2012. Besides preserving an invaluable part of history, this collaborative effort also serves as a model for future endeavors in archiving "at risk" web sources in the field of East Asian Studies and beyond. The project received enthusiastic endorsement from scholars and library administrators. It was funded over a two-year period (July 1, 2015 to June 30, 2017) by the Mellon Foundation-CEAL Innovation Grant Program. At the close of the grant term, we are proud to have accomplished all the major goals we set out in the initial proposal, including:

- Developing software with the capability to capture and store Sina Weibo social media content, and able to capture and export social media data in other East Asian languages;
- Building a collection of Chinese web content and social media data with finding aids and metadata to enable discovery via the library catalog;
- Documenting the project’s technical implementation, technical challenges and solutions to help future similar projects;
- Presenting and promoting the collections at national professional venues such as the annual conferences of CEAL (March, 2016) and ACRL (March, 2017).
Below is the final report of the project. We will start with a backgrounder on Chinese social media and the anti-corruption campaign, a description of our general approach and outcome, followed by detailed reports on two components of the project: the Archive-It collection at JHU and Social Feed Manager collection at GW.

1. Chinese Social Media and the Anti-corruption Campaign: the Project Rationale
The project focused on archiving online postings related to the Chinese anti-corruption campaign on Chinese social media platforms such as Sina Weibo, Sina blog and other blogging and news sites. Before we go into the details of the project, we would like to give a brief summary of our research on Chinese social media and the anti-corruption campaign that led to the idea and eventual proposal and implementation of the project.

1.1. Why social media?
The past decade has witnessed an explosion of social media, thanks to the rapid advancement of the internet and mobile technology. Social media have become an integral part of our life today. Yet, because of the transient, ephemeral nature of social media postings, these born digital materials are prone to disappearing quickly, becoming lost forever to future researchers. For this reason, how to preserve social media records for future studies has presented a challenge to archivists and librarians.

In the case of China, preserving social media records has become even more critical. The ruling Communist Party of China (CPC) practices censorship of all forms of media for a variety of perceived reasons, ranging from political to moral to economic. Traditional media such as radio, television and newspapers are under tight government control and serve primarily as an extension of official Party communication. Under such conditions, social media has emerged as a significant public space where ordinary Chinese citizens can express their unedited views, including criticisms of the government. Access to Facebook, Twitter, YouTube and other major social media platforms is now blocked by the Chinese government; however, WeChat instant messaging, Sina Weibo — a microblogging platform similar to Twitter — and other local platforms offer the Chinese people local choices to meet their social media needs. Blogging and microblogging play a crucial role in China’s socio-political activism. They represent non-official voices of history that often challenge the official narratives in government-run, tightly-controlled traditional media.

In addition to having the world’s largest Internet user base — nearly 600 million people, more than double the 250 million users in the United States — China also has the world’s most active environment for social media, with 100 million bloggers and 300 million microbloggers at present.¹ Despite the government’s continual efforts to monitor and

control Internet traffic, the sheer volume and scale of Chinese social media makes it technically impossible to censor everything. As a result, the Internet, and social media in particular, has “effectively eliminated the government’s monopoly on information.”

An indicator of its lasting significance, over the last several years Chinese social media has become a favorite subject for scholarship by Chinese studies scholars worldwide. A search on “social media and China” in ProQuest’s Dissertations and Theses produces 10,216 hits, with the majority of titles (9,926) published since 2010. A more focused search on Weibo (微博) in the Chinese Dissertations and Theses database (by Wanfang Data) results in an even more hits: 12,482, the majority of which have been published since 2011.

Initiated in July 2015, this web archiving project marks a first step towards archiving Chinese social media. With joint resources and technical expertise, librarians from GWU, JHU and GU have collaboratively developed tools and strategies to archive selected Chinese social media sites. The project immediately benefits scholars interested in the use of social media as a tool for Chinese socio-political activism and expression within civil society.

Many scholars at the three collaborating universities have recognized the importance of the work of this project. For example, Joel Andreas, Associate Professor of Sociology at Johns Hopkins University, upon learning of the project, shared his manuscript paper just recently submitted for publication with the project team. Entitled “Mass Supervision and the Bureaucratization of Governance in China,” the article studies cyber-activism as a new form of “mass supervision” that was once popular during the Maoist era (1949-1976). Similarly, Jackson Woods, a Ph.D. student in Political Science at George Washington University, while studying political discourse on the web in contemporary China, was excited to be able to utilize the collection in his dissertation project.

In addition to benefiting researchers, this project also creates tools and documents methods that librarians who wish to pursue similar projects can use to archive Chinese blogs and microblogs. One of the project outcomes is a companion paper documenting the web archiving process and lessons learned from the experience, with particular emphasis on the unique challenges encountered when collecting China-based social media content. Given the scarcity of such information, this document will fill a major gap and likely be of significant value to the wider library community.

1.2. Why the Anti-corruption Campaign?
This project focuses on China’s ongoing anti-corruption campaign. Even though fighting corruption has been a perennial theme throughout Chinese history, the Anti-Corruption Campaign launched by the Chinese President Xi Jinping right after he was named General

---

Secretary of the CPC in November 2012 has evolved into a major political movement generating great interest from scholars both inside and outside China. More significantly, social media has played a crucial role in the ongoing campaign.

What distinguishes China’s anti-corruption legal framework from other nations’ is that there is not a specific anti-corruption law in China. More generally, the criminal law of China serves as the leading legal weapon to define and punish corrupt behaviors. According to China’s Criminal Statutory Law, corruption is generally viewed as any activity which employs public power to earn private benefits. This is a very loose definition of corruption in the Chinese setting but is widely accepted by the Chinese government.

Corruption among China’s government servants was historically believed to be the major cause of dynastic collapse, including the collapse of the last rulers of China before the communists, the KMT Nationalists, who left for Taiwan in 1949. After Mao Zedong’s rise to power in 1949 and until his death in 1976, a series of political campaigns took aim at cadre corruption, reaching a climax during the decade of the Cultural Revolution (1966-1976).

The next fifteen years from 1976 to 1992 saw China open its doors during the Deng Xiaoping era. The wide-ranging reforms central to this economic transformation brought corruption in China to new heights. Corrupt practices allowed China’s top leaders to profit handsomely as they engaged in various business enterprises. Their roles as middlemen or brokers (Guandao 官倒 bureaucrat-profiteers) became the major cause of China’s pro-democracy movement in 1989.

Under the leadership of Jiang Zemin, Deng’s successor, the decade from 1992 to 2002 witnessed a period of increased emphasis on anti-corruption. In 1993, the term “anti-corruption” officially entered into government documents and the Party Charter. Although the central government took heightened actions to crack down on corruption, the absence of an effective legal framework hindered results. Additionally, ineffective laws and regulations permitted top corporate officers of state-owned enterprises to avoid repercussions for a wide range of corrupt practices.

The first decade of the 21st century saw the advent of more institutional means to curb

---


corruption. According to the National Bureau of Corruption Prevention, more than 43,000 officials were put on trial between 2003 and 2011. In 2013, under Xi Jinping’s new leadership, the CPC launched a new, more effective “war on corruption.” To date, the Central Discipline and Inspection Committee (CDIC) led by one of Xi’s most trusted colleagues, Wang Qishan, has announced the investigation of hundreds of high-ranking officials commonly known as “tigers (大老虎)” and hundreds of thousands of low-ranking officials commonly known as “flies (小苍蝇).” The tigers who were prosecuted and sentenced in 2014 and 2015 included China’s top leaders including Bo Xilai (薄熙来), Zhou Yongkang (周永康), Xu Caihou (徐才厚) and Ling Jihua (令计划), known as the new Gang of Four.

Even though countering corruption has become the government’s central objective, it is still not legally defined. Some critics have claimed that the current anti-corruption campaign is the result of a power struggle to cleanse and purge political opponents. Such discussions attracted wide attention on social media. While the central government has depended on reliable propaganda tools including newspapers, TV, magazines and other traditional media, they also engage new technologies such as social media to promote and defend the campaign. Government communication via social media has increased over the last five years to become a very important informational and interactive tool to combat corruption. Popular social media outlets include WeChat, blog, micro-blog, BBS (Bulletin-board Sites), forums and online news. At the same time, since most Chinese citizens can access the Internet, spreading evidence of suspected corruption can create huge social pressure to force government agencies to intervene.

The advantages of using social media over traditional media for Chinese citizens wanting to communicate are manifold: it is more open and accessible; information spreads more quickly; the costs are much lower; and contributions can be initiated more anonymously. As a result, individuals are able to effectively expose corrupt practices or critique government action that would not otherwise be revealed, and can do so with less worry about retaliation. Social media plays a significant role in empowering individuals who otherwise would not have adequate protections, either legally or through the CPC’s practice in ruling the nation. The government in turns blocks access to certain sites, strengthens firewalls and launches other strict controls on social media in order to reestablish control over anti-corruption initiatives.

Based on the historical significance of the event, scholarly interest, and the crucial role of social media, we decided that the Anti-Corruption Campaign would be a perfect topic for our collective efforts. Our project aims to archive both the government’s Internet-based/social-media accounts and private accounts focused on the anti-corruption campaign due to the concern that some social media contents might forever be “lost” due to government censorship and the passage of time. Despite its clear historical and scholarly importance,
Archiving social media has remained a major challenge for librarians and archivists worldwide. The Internet Archive and the Twitter Archive Project at the Library of Congress represent two major steps in the West to archive web content. However, in the case of Chinese Internet sources, especially Chinese social media, little has been done to preserve this valuable source of Chinese history, either inside or outside China.6

For this reason, we view the project as an opportunity to contribute to the field of China Studies as well as librarianship and web archiving. In addition to preserving an invaluable part of Chinese history, this collaborative project will also serve as a model for future endeavors in archiving “at risk” web sources, and will create software for collecting Weibo content that will be made available to others with an interest in creating Weibo collections in the future.

2. General Approach and Outcome: a Summary
This project focuses on identifying and archiving content from two sources: postings on “static” blogging and news sites on the Internet, and microblogging from “dynamic” social media sites such as Sina Weibo.

For the “static” blogging sites, we use Archive It (https://archive-it.org/) from the Internet Archive (https://archive.org/) to collect and preserve selected records. Archive-It offers a convenient interface for librarians and archivists to build collections by identifying sites to crawl and then specifying a crawling schedule; curators also add collection metadata such as subject tags, to make the collection more discoverable. This project’s Archive-It collection, curated by Johns Hopkins University Libraries, has so far archived 1,300 Chinese blogs entries and news articles related to the Anti-Corruption campaign. The collection is publicly viewable at https://archive-it.org/collections/6314.

Collecting “dynamic” social media content can be viewed as a type of web archiving, but it has some distinguishing features that call for more specialized software tools.7 Social Feed Manager (SFM), an open-source software developed by the George Washington University Libraries with the support of grants from the Institute for Museum and Library Services and the National Historical Preservation and Records Commission, is designed specifically to enable archivists and researchers to build collections of social media data.8 SFM provides

---


8 Further information about the Social Feed Manager project and software application can be found at https://gwu-libraries.github.io/sfm-ui/.
the ability to collect content from Twitter, Tumblr, and Flickr; this project’s grant from CEAL supported the GW’s development of a new Sina Weibo harvester for SFM to enable this and potential future Weibo archiving efforts. This will be discussed in more detail later in this paper.

In addition to the collection of data, our project has also accomplished the following goals:

2.1. Being the first collaborative effort on the institutional level to archive Chinese social media. With expertise in China studies librarianship and in social media web archiving technology from three top research universities, we are in a unique and advantageous position to take on the challenge of Chinese social media archiving.

The project benefits from strong faculty support in China studies from all three institutions. Professor Erin Chung, chair of East Asian Studies Department at Johns Hopkins University, as well as other core faculty members in the department such as Professor Lingxing Hao and Joel Andreas, all have pledged their full support and agreed to serve on the Advisory Committee. Core faculty at Georgetown University, such as Professor Jingyuan Zhang, Chair of the Department of East Asian Languages and Cultures, Professor Philip Kafalas of the same department and several PhD students in Georgetown’s departments of History and Sociology all expressed strong enthusiasm and interest in the project. China Studies faculty at George Washington University are also strongly supportive, including Bruce Dickson, Chair of the Political Science Department and a specialist on Chinese politics, authoritarian regimes, regime change, and state-society relations.

The expertise of George Washington University Libraries in social media archiving is unique and has been essential to creating a Weibo archiving capability; GW also contributes experience using Archive-It to build web archive collections. GW Libraries is a national leader in the area of social media archiving and has several years of experience using its locally-developed software, Social Feed Manager, to build collections for researchers mainly from Twitter, which is similar to Weibo in key respects. GW earned the support of two major grants: an IMLS Sparks! Ignition Grant (2013)\(^9\) to develop the initial SFM prototype, and an innovation grant from National Historical Publications and Records Commission (2014),\(^10\) to extend SFM to automate capturing and preserving data from Twitter, Tumblr, Flickr, and now Weibo, to build a user interface for SFM that enables researchers to build collections directly, and to re-architect the application to make it easy for other institutions to set up. Daniel Chudnov, former manager of the Twitter Project at the Library of Congress, and Principal Investigator of GWU Libraries social media archiving grants through 2015, helped

\(^{9}\) Institute of Museum and Library Services grant #LG-46-13-0257-13.
\(^{10}\) National Historical Publications and Records Commission grant #NARDI-14-50017-14.
envision the project. Daniel Kerchner, Senior Software Developer, the technical lead on the project and who has contributed to the SFM code, serves as a consultant to GW researchers building collections with SFM to analyze social media content in the contexts of a variety of academic fields and has participated in web and social media preservation collaborations at a national level.

2.2. Developing open access software usable by librarians and researchers that is adapted to archive social media content from Sina Weibo.11 Archiving social media content from any platform generally requires retrieving data using the platform’s Application Programming Interface (API). Each platform’s API is different, and each provides varying levels of access to the data. Only institutions with special arrangements, such as in the case of the Library of Congress Twitter Archive Project, have access to all content from a social media platform.12 Even in the rare cases when all of the platform’s content can be obtained, the volume of data presents a formidable challenge in storing and providing access to the data. This challenge has yet to be satisfactorily resolved.

For more realistic scenarios, social media APIs provide a service to make more focused requests for data. These APIs generally provide the ability for an application to request the content of specific public microblogs (accounts), to request posts mentioning specific terms (such as hashtags, names, or keywords), or to request a small sample of all posts; other types of requests are often available as well. As SFM has been developed to implement these interactions to retrieve data from Twitter and other platforms, SFM was a natural choice to extend in a similar manner to retrieve social media data using Weibo’s API. Furthermore, SFM fills a unique niche as open-source software for building social media collections, in contrast to commercial options for obtaining this data. More specifically, we are not aware of robust open-source software for collecting data from Weibo; as such, SFM enhanced with Weibo collection capability provides an important new option indeed, particularly for librarians and researchers.

2.3. Integrating the records for Chinese social media collections into the existing library discovery tools such as LibGuides and online catalog.

In order to facilitate the use of the archived social media collections, we designed a LibGuide page dedicated to the project (http://libguides.gwu.edu/c.php?g=637999&p=4466423). The guide introduces the user to the collections, including an overview of the project, sample data, as well as link to collected data (in the case of Archive It). At Johns Hopkins, the Archive It collection is also discoverable through the library catalog, with links leading directly to the

11 Please see the companion paper for technical specifications and other details.
online archived data (https://catalyst.library.jhu.edu/catalog/bib_6635884).

Because of legal and privacy concerns, it is more challenging to provide access to the Weibo data; the company specifically forbids republishing their data. After investigating the Library of Congress Twitter Archiving Project and consulting with GW Libraries staff responsible for the Social Feed Manager, we decided that providing open access to the collected data, while ideal, is out of scope of this project and would involve legal and ethical challenges\textsuperscript{13} that cannot be resolved during the grant term. As a result, we will rely on the LibGuide and other publicity channels to promote and facilitate the use of the collection.

II. ARCHIVE-IT AND CHINESE BLOG COLLECTION

As mentioned above, our web archiving project consists of two components. One is to collect data from Sina Weibo using Social Feed Manager; this is conducted by the team at George Washington University. The other is to collect data from various blogging and news sites using Archive It and is done by the team at Johns Hopkins. Below is a summary of building the Chinese social media collection using Archive It at JHU.

1. Why Archive-It?

Launched by the Internet Archive in 2006, Archive-It is a subscription-based web archiving service that allows organizations to collect and preserve their digital content. According to the “Learn More” page on their website, Archive-It currently serves over 400 partner organizations including college and university libraries, state libraries, museums and art libraries, federal institutions and NGOs, and public libraries.\textsuperscript{14}

Archive-It allows partners to capture and manage web published materials including websites, blogs, and social media sites. It could be used to capture postings on “dynamic” sites such as Twitter, but it seems to work best when used to capture postings on “static sites.” Compared with the public Wayback Machine, a free service provided by the Internet Archive, the subscription based Archive-It brings many important benefits. For example, the public Wayback Machine is just one huge collection holding all archived materials, the user having very little control over the collected data, whereas Archive-It allows the partner to build their own focused collections based on themes and types of contents. Archive-It gives full control over the scope of the collection and frequency. It comes with full-text search functionality, content cataloged with metadata, and complete access control. The archived data are duplicated with copies residing at Internet Archive data centers on multiple servers. The service works with partners who may wish to have redundant copies for preservation.

\textsuperscript{13} A short list of recent publications on the ethical and legal considerations in collecting and using social media data can be found at https://gwu-libraries.github.io/sfm-ui/resources/ethics, accessed January 31, 2017.

\textsuperscript{14} Archive-It, “Learn More”, retrieved at https://archive-it.org/learn-more/.
More importantly, the technical support for Archive-It is excellent and crucial, especially for those who just start out.

In short, because it offers a package solution to our collecting, cataloging, access, and preservation needs, we chose Archive-It as the tool of choice for archiving Chinese blogs.

2. Using Archive-It: Issues and Solutions
Anyone can learn and use Archive-It; no programming experience is required. That said, it does have a learning curve and needs some training and practice in order to use it effectively. For prospective partners, the Archive-It hosts twice-monthly webinars that provide an overview and demonstrations of the product. The webinar consists of a PowerPoint presentation that thoroughly discusses the features, customizable options of the product, archiving needs of the trainee, as well as some sample work done by other institutions. We would highly recommend starting with the training webinar for anyone unfamiliar with the tool.

Next comes the experimentation phase. It took us about three weeks to thoroughly familiarize ourselves with the tool. Eventually we came up with a three-step process that worked fine for us.

Step One: Identify the target material.
Finding a target blog to archive seems an easy step. However, it became an issue for us almost as soon as we started. The Archive-It is designed to crawl a specific website or page (via its URL, or “seed” as it is called in the Archive-It), at a frequency designated by the user. For our project, since we only collect postings related to the theme of “anti-corruption”, it became obvious to us that we could not use the default approach, crawling the same site or page and archiving whatever may appear on the page at the moment. Instead, we opted for a “curation” approach, searching and finding the relevant content and then archiving that article only. Fortunately, the Archive-It does give the option to allow us to just crawl one page at a time and save that page.

In order to find relevant content to archive, we built a Google Blog Search Engine that searches over 50 major blogging sites and news sites in China. We also came up with a list of keywords on various themes of the anti-corruption campaign as well as names of key people and high profile cases. Every week, we conduct the same searches via the special search engine and identify target articles for archiving. The graduate assistant who worked on the project was a Ph.D. student in political science specialized in contemporary Chinese politics. So we have high confidence in the quality of data collected.

Step Two: Collecting Data
This probably is the most challenging part for people who never used the tool before. It will
be confusing and frustrating at times. For example, even if we find the right content and feed the “seed” (URL) to the Archive-It, the Archive-It may or may not like it. Not all crawls will be successful. Also, the Archive-It automatically collects all the “documents” including any advertisements, extraneous links and images that happen to be on the page. If not careful, you may quickly use up the storage capacity that you have purchased. We would recommend starting with one page at a time, and always use “test crawl” first before permanently archiving a file.

After you get to know the process well and feel comfortable using it, you may increase crawling up to three or five “seeds” at time. As the collect grows, it is important to keep track of the collected documents and avoid duplication. That is because the production mode of Archive-It only gives you a record number and partial URL for each document, and it is difficult to know which record corresponds to which document. Once the collection grows to more than 100, it becomes increasingly difficult to remember whether a certain blog has already been archived or not. For this reason, we also recommend using a word document or spreadsheet to keep track of the entire document URLs and their titles, dates of their crawling, and whether the crawl was successful.

Step Three: Creating Metadata

One nice feature of Archive-It is the ability to create our own metadata specific to the collection that we are trying to build. For each record collected, there is a meta form to tag the record. These metadata will then be used on the public site to facilitate browsing. We decided that we would use “subject,” “date,” “type,” “language,” “collector,” and “source” as our main categories. Under “subject,” we used various descriptors that pertain to the data we are collecting. For example, we knew we wanted to collect data from both the national level and local levels; we wanted our data to cover both high profile cases (so-called “tigers”) and smaller cases (“flies”); and we wanted our collection to focus on government corruption, but also include cases from major industries such as banking, energy, education and so on. Accordingly, we came up with subject descriptors that reflected the target contents that we tried to collect. In the end, when browsing our collection online, one can either search for the content they are interested in, or click on the corresponding subject filters on the menu bar and easily access the materials.

3. Outcome

The official project start date was July 1, 2015. It happened to be in the summer and most students were out. Recruiting a qualified student assistant proved to be difficult. We did not hire a student assistant until October 2015, three months after the project kicked off. Because of the delay, plus time for training, we were only able to collect about 300 records in our collection. Then in the second year, we were in full production mode and added about 1,000 more entries to our collection. In addition to the metadata, each entry also has both
Chinese title and its English translation, and a brief description in English of the content. The entire collection is full text searchable, and easily browsable through subject filters. Below is a (partial) screenshot of the first page of the collection.

To facilitate discovery and access, we created a catalog record and added to the JHU library catalog (See below). Anyone who searches the online catalog with keywords such as “China social media” or “China corruption” will be able to find the record and link to the collection.
Although the project term comes to an end, because JHU has an institutional subscription to the Archive-It, we will continue to monitor the anti-corruption campaign in China and add new entries to the collection in the years to come.

III. SOCIAL FEED MANAGER (SFM) AND SINA WEIBO COLLECTION

Collecting social media data from social media platforms such as Twitter and Sina Weibo are governed by the affordances of the Application Programming Interfaces (APIs). In the case of Sina Weibo, those affordances are quite limited. The Sina Weibo API supports two methods that are relevant for the topical collecting of Weibo posts: the “friends timeline” method which is available from the basic API and the “topic search” method which is available from the advanced API. The “friends timeline” returns a subset of the Weibo entries posted by a user and the other users followed by the user. The “topic search” method returns a subset of tweets matching a topic. A topic is a single keyword bracketed by “#” in the text of a Weibo, allowing users to generate a folksonomy on the Sina Weibo platform, similar to the role of hashtags on the Twitter platform. (The API methods are explained in greater details in a companion piece to this article. The companion piece covers API-based social media collecting and the Sina Weibo platform and API in greater detail.)

1. Collecting social media data with SFM
Step One: Identify the target material.

Like collecting blogs, the first step in collecting data from Sina Weibo is to identify target material. For a collection built from the “friends timeline” method this involves identifying a list of users that post about a relevant theme. For a collection built from the “topic search” method this involves identifying a list of topics that are used to mark Weibo posts that are relevant to the theme.

In both these cases, the most straightforward way to identify users or topics is via the Sina Weibo website (http://Weibo.com). First, in order to view full content of the Sina Weibo website, one must be a registered user. Second, the selection of accounts is based on their role, accountability and relevancy to a specific topic. We generally follow three types of accounts: government accounts, media accounts and private accounts. Government accounts and media accounts with a label of “官方微博 Official Weibo” indicate that Sina Weibo has authenticated their identity. A private account with “微博认证 Weibo certification)” or a label of V(IP) could prove its accountability. In terms of identifying accounts with a focus on Anti-Corruption Campaign, government and media accounts post about diverse topics, of which the anti-corruption campaign may be only one of the topics. For examples, 澎湃新闻 is a newly created online news site which offers timely and relatively objective reports; one of its foci is exposing corrupt officials and their cases. Private accounts, however, must be highly focused on the anti-corruption campaign and have a respectable number of active followers. In the meantime, private accounts also play an important role in revealing potential corruption cases and individuals. Third, identifying these accounts requires professional knowledge in contemporary China’s politics and familiarity with Chinese media, government, and Sina Weibo platforms. Our China Studies librarians who possess the knowledge and skills first identified the most well-known government and media accounts. The second step is to identify more accounts based on the identified accounts’ posts. The most effective way is to check most authoritative accounts’ Weibo. If an account is tagged or its Weibo is being reposted by the authoritative accounts, it could be a potential candidate. In terms of identifying topics, some of them were found from hashtags that appeared in the identified accounts’ Weibo. The others need librarians’ knowledge on Anti-Corruption through reading scholarly publications and news.

For the project, initial work identifying topics and users relevant to the anti-corruption campaign in China was performed by Jing Zhong (former China Studies Librarian, George Washington University) and updated by her successor, Yan He. Thirty-two users were identified. (For privacy reasons, the list of users will not be disclosed here.) Forty-five topics were identified. These include 中央电视台, 举报, 党内监督, and 党纪国法. For various reasons, we were not able to collect all of the users and topics, e.g., some accounts were closed and some topics were no longer being used.
Step 2: Create a Sina Weibo account and get API credentials

Prior to beginning collecting Weibos, using the “friends timeline” method requires creating a “dummy” Sina Weibo account that is to be used exclusively for collecting. That account should follow only the users from which you want to collect Weibos. Using a personal account is not recommended.

Like other social media platforms, Sina Weibo requires applications such as SFM to have credentials to access the API. In turn, SFM requires users to acquire those credentials and supply them to SFM. Since the process of getting credentials from Sina Weibo can be confusing, we have documented the process to reduce the barrier for users at other institutions. Fortunately, once the credentials are acquired, the process of providing those credentials to SFM is via a web form and is a simple process.

Step 3: Create a collection

SFM is designed to allow users to directly create and administer their own collections without mediation. Thus, to create a collection, a user completes a web form supplying the name of the collection, a description of the collection, a collecting schedule, and some other options.

---

For a topics collection, the user must also provide the topic. (Like web harvesting, these are referred to as “seeds” in SFM.) The collection can then be turned on, which initiates harvesting in accordance with the schedule.

Here is part of the page showing the anti-corruption campaign timeline collection on GW’s SFM instance:
Step 4: Wait

SFM will repeatedly perform harvests in accordance with the schedule. In the case the anti-corruption campaign timeline collection, this is quite often -- every 30 minutes -- to try to
collect as many Weibo posts as possible. As SFM is proceeding with harvesting, users can review progress. Below is a list of harvests:

<table>
<thead>
<tr>
<th>Type</th>
<th>Requested</th>
<th>Updated/Completed</th>
<th>Status</th>
<th>Stats</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 10:23:51 a.m. EDT</td>
<td>May 12, 2017, 10:24:00 a.m. EDT</td>
<td>Success</td>
<td>24 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 9:23:51 a.m. EDT</td>
<td>May 12, 2017, 9:24:00 a.m. EDT</td>
<td>Success</td>
<td>27 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 8:23:51 a.m. EDT</td>
<td>May 12, 2017, 8:24:00 a.m. EDT</td>
<td>Success</td>
<td>23 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 7:23:51 a.m. EDT</td>
<td>May 12, 2017, 7:24:00 a.m. EDT</td>
<td>Success</td>
<td>25 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 6:23:51 a.m. EDT</td>
<td>May 12, 2017, 6:24:00 a.m. EDT</td>
<td>Success</td>
<td>51 weibos</td>
<td></td>
</tr>
</tbody>
</table>

and here is a diagram of the number of Weibo posts collected over the last month:

Step 5: Adjust users and topics

Sina Weibo is a very dynamic environment. To account for this, Sina Weibo should be monitored for new users to follow or topics to collect. (The reverse is true as well: it may be desirable to remove users or topics.) Adding a user is done by following the user from the
dummy" account on the Sina Weibo website. To add a topic, a new topic collection is added in SFM.

2. Using Weibo data collected with SFM
There are a number of mechanisms for accessing social media data collected by SFM. However, by far the easiest and most common is to export it to a spreadsheet.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
<th>N</th>
<th>O</th>
<th>P</th>
<th>Q</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>created_at</td>
<td>screen_name</td>
<td>followers_count</td>
<td>friends_count</td>
<td>reposts_count</td>
<td>topics</td>
<td>in_reply_to_status_id</td>
<td>retweet_count</td>
<td>favorite_count</td>
<td>retweeted_status</td>
<td>in_reply_to_status_id</td>
<td>retweet_count</td>
<td>favorite_count</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Spreadsheets are a flexible way of sharing with a scholar, as the Weibo data can be directly viewed or loaded into analytic software.

3. Comparison to web archiving with Archive-It
It is worth drawing attention to some of the salient differences between collecting web resources such as blogs or articles and collecting Weibo posts from Sina Weibo, as they have slightly different natures. First, collecting of Weibo posts is possible for current posts only; the ability to get historical Weibo posts is extremely limited. Web archiving, on the other hand, has a much larger collection window. While the life of a web page is notoriously short, it is measured in months not the minutes in which a Weibo can be collected from the API.

Second, collecting of social media data is done on an ongoing basis. For example, once a collection is started for a topic, that harvest is repeated over a period of time. By comparison, web harvesting of blogs or articles is generally a one-time activity. The curator of a web harvesting collection must regularly be looking for new blogs or articles to harvest. (One interesting approach that was not explored in this project would be to attempt to identify new blogs or articles from the links contained in harvested Weibo posts).

4. Outcome
The SFM and associated Weibo harvester work have made a number of contributions that we believe move the field forward in significant ways.

First, the SFM Weibo harvester is, to our knowledge, the first of its kind; the enhancements to SFM and creation of the Sina Weibo harvester were developed and documented with a broader goal of ensuring that they can be set up and run by other cultural heritage
institutions interested in building collections of Sina Weibo data. After much work resolving a series of technical barriers and other impediments, GW’s SFM team first released the Sina Weibo harvester on May 31, 2016. Initially it only supported the “friends timeline” method, with support for the “topic search” method added on March 14, 2017. The SFM and Weibo harvester code is public and open-source, so that others looking to utilize and build on it have full transparency into its methodology, and can more easily build upon work to date.

Secondly, we have publicly documented a number of lessons learned in using the Weibo API for this type of work to help guide others seeking to embark on related work. This includes our creation of instructions for obtaining Weibo API credentials, and our writings on some of the challenges in working with the Weibo API and our solutions to those challenges.

Finally, we have used both of the above to successfully create the first Sina Weibo collection. The Sina Weibo collection was started on June 29, 2016. Since then, over 110,000 Weibo posts have been collected across all collections. We look forward to scholarly interest in analyzing the Weibo posts.

IV CONCLUSION

Since the start of the project in 2014, we have received strong support from library colleagues and administrators, faculty and graduate students from the three participating universities and beyond. Over the past two years, the collaborators from the three institutions, including East Asian Studies librarians, IT and administrative personnel, met multiple times to discuss the planning and implementation of the project. Despite extra work being added to our regular workload, all three librarian collaborators welcomed the opportunity to venture outside the realm of traditional librarianship and learn and use new technology to help develop new types of collections. And our effort has paid off. We have managed to build a digital collection of millions of Weibo posts and over a thousand blog articles, and helped to preserve a valuable part of Chinese history. In addition, we hope our project will inspire more similar projects in the field of East Asian Studies librarianship and beyond, to collect and preserve ephemeral online materials for future studies.
Collecting Social Media Data from the Sina Weibo Api

Justin Littman
Daniel Kerchner
Yan He
Yecheng Tan
Cathy Zeljak

Abstract
As part of a larger collaborative project to preserve non-official voices in China’s anti-corruption campaign, a team from George Washington University Libraries was tasked with collecting content from Sina Weibo, a massively popular Chinese social media site. Based on their experience developing Social Feed Manager (SFM), an open-source tool for collecting social media data to support research and constructing archives, the team added support for Sina Weibo. This paper provides introductions to both the Sina Weibo platform and the approach of collecting social media from an Application Programming Interface (API). Further, it describes SFM’s support for Sina Weibo, the challenges in working with the Sina Weibo API, guidance for other institutions interested in conducting similar work, and a brief characterization of the content collected about China’s anti-corruption campaign.

Introduction
Sina Weibo1 is a massively popular Chinese social media site, in use by over 100 million users per day (Freier 2015). Given its prominent role in communicating the “everyday voices” of the Chinese public, the value of collecting content from Sina Weibo was self-evident when a team from Johns Hopkins University, George Washington (GW) University, and Georgetown University initiated a collaborative project to preserve non-official voices in China’s anti-corruption campaign. (For background on and motivation for the grant, see the companion report, “Web-archiving Chinese Social Media: Final Project Report.”) Awarded a Mellon innovation grant administered by the Council on East Asian Libraries,2 this team took a two-prong approach:

1. Collect blog content using Archive-It, a service provided by the Internet Archive for performing web archiving.
2. Enhance Social Feed Manager (SFM), an open-source tool developed at GW Libraries, to collect weibos (posts) from the application programming interface (API) of Sina Weibo.

---

1 http://weibo.com/
2 http://www.eastasianlib.org/MellonGrants.htm
The goal of this paper is to explore the second prong, the enhancement of Social Feed Manager, by:

1. Giving a brief overview of Sina Weibo.
2. Explaining API-based social media collecting.
3. Describing the affordances of the Weibo API.
4. Providing an overview of SFM and the Weibo social media harvester.
5. Document some challenges in working with the Weibo API.
6. Characterize the anti-corruption campaign-related social media data that has been collected to date.

**Sina Weibo**

Sina Weibo is a micro-blogging social media platform similar to Twitter. The core of Sina Weibo is the “weibo” (“微博”, meaning microblog), a post analogous to a tweet. Users on Sina Weibo perform two main functions with weibos: they author weibos and they read weibos.

When authoring a weibo, users are restricted to 2000 Chinese characters, an increase from the earlier limitation of 140 Chinese characters, which mirrored Twitter’s character limit. However, when viewing a weibo, only the first 140 Chinese characters are displayed, with a link allowing the rest of the text to be viewed. To mention another user, the user’s username is preceded by an “@”. To mark a term in a weibo as a topic, it is bracketed by “#” (similar to Twitter’s hashtag -- more on this below). Weibos can include URLs; embedded images or multimedia items; and/or emoticons.

**Sina Weibo**

The following image shows a screenshot of a weibo post on Sina Weibo. The screenshot includes options for adding various types of content to the weibo, such as emojis, photos, videos, topics, articles, and the ability to make the content public or private. The interface is user-friendly, allowing users to easily share their thoughts and interact with others on the platform.

**What’s new with you?**

 diversos temas relacionados a la política anti-corrupción. Estas publicaciones han sido recopiladas hasta la fecha.

**Sina Weibo**

La plataforma de microblogging Sina Weibo es similar a Twitter. El núcleo de Sina Weibo es el “weibo” (“微博”, que significa microblog), similar a un tweet. Los usuarios en Sina Weibo realizan dos funciones principales con weibos: escribir weibos y leerlos.

Al escribir un weibo, los usuarios están limitados a 2000 caracteres chinos, un aumento respecto a la limitación anterior de 140 caracteres chinos, que reflejaba el límite de caracteres de Twitter. Sin embargo, al leer un weibo, solo se muestran los primeros 140 caracteres chinos, con un enlace que permite ver el resto del texto. Para mencionar a otro usuario, el nombre de usuario del usuario precede un “@”. Para marcar un término en un weibo como un tema, se lo bracketa con “#” (similar a Twitter’s hashtag -- más en lo que sigue). Los weibos pueden incluir URLs; imágenes o elementos multimedia; y/o emoticonos.

**Sina Weibo**

La siguiente imagen muestra una captura de pantalla de un post de weibo en Sina Weibo. La captura de pantalla incluye opciones para agregar varios tipos de contenido al weibo, como emojis, fotos, videos, temas, artículos y la capacidad de hacer el contenido público o privado. La interfaz es amigable, permitiendo a los usuarios fácilmente compartir sus pensamientos e interactuar con otros en la plataforma.

**What’s new with you?**

En resumen, este artículo tiene como objetivo explorar el segundo punto, la mejora de Social Feed Manager, mediante:

1. Dar un breve resumen de Sina Weibo.
2. Explicar la recopilación de medios sociales basada en API.
3. Describir las prestaciones del API de Weibo.
4. Dar una visión general de SFM y el recolector de medios sociales de Weibo.
5. Documentar algunos desafíos en el trabajo con el API de Weibo.
6. Caracterizar los datos de medios sociales relacionados con la campaña contra la corrupción que se han recopilado hasta la fecha.

**Sina Weibo**

Sina Weibo es una plataforma de microblogging similar a Twitter. El núcleo de Sina Weibo es el “weibo” (“微博”, que significa microblog), similar a un tweet. Los usuarios en Sina Weibo realizan dos funciones principales con weibos: escribir weibos y leerlos.

Al escribir un weibo, los usuarios están limitados a 2000 caracteres chinos, un aumento respecto a la limitación anterior de 140 caracteres chinos, que reflejaba el límite de caracteres de Twitter. Sin embargo, al leer un weibo, solo se muestran los primeros 140 caracteres chinos, con un enlace que permite ver el resto del texto. Para mencionar a otro usuario, el nombre de usuario del usuario precede un “@”. Para marcar un término en un weibo como un tema, se lo bracketa con “#” (similar a Twitter’s hashtag -- más en lo que sigue). Los weibos pueden incluir URLs; imágenes o elementos multimedia; y/o emoticonos.
The primary mechanism for reading weibos is a user’s timeline. A timeline is a list of weibos that is primarily determined by the user accounts that the user follows. A user’s timeline contains the weibos posted by the followed user accounts. Upon reading a weibo, a user can comment on it or repost the weibo, similar to retweeting on Twitter.

An example user page would appear as follows:

Like other social media platforms, Sina Weibo can be accessed from a website or smartphone application (i.e., mobile). Sina Weibo has other social features, but those described above are the most relevant for this discussion.

API-based social media collecting
When a user goes to a website of a social media platform, e.g., http://us.weibo.com/gb, she interacts with the web interface of the platform. This interface is intended to facilitate the interaction between a human user (using a web browser) and the social media platform. However, many social media platforms provide an additional interface, the API, to facilitate the interaction between software and the social media platform. So, for example, when the Twitter app on your phone needs to fetch your tweets, it requests those tweets from Twitter’s API. Unlike the web interface, which returns HTML that is optimized for rendering in a browser, the API returns structured data, usually in a format called JSON, which is optimized for processing by software.3

Both the web interface and the API can be used to collect content from a social media platform. Collecting from the HTML/web interface requires using web harvesting software such as Heritrix4 or WebRecorder;5 collecting from an API requires software such as SFM or Twarc.6 Social media harvesting and web harvesting can be considered complementary, with different strengths and weaknesses (Littman et al. 2016). Some of the strengths of collecting from the API include:

● Data collected by the API is in a structured format such as JSON or XML. These formats are amenable to research that involves applying computational techniques. (This will be more evident when we look at the anatomy of a weibo below.) Extracting similar data from web pages would be extremely difficult.

● Some social media APIs provide metadata that is not available from the website. For example, the Twitter API provides the name of the device/application used to author a tweet.

● Social media platforms tend to keep their APIs stable and when making a change will generally announce it in advance. This is because there are potentially numerous other application (e.g., smartphone apps) that depend on the API. In contrast, social media web pages regularly change in small ways that make extracting data problematic.

● To optimize the user experience, social media websites make heavy use of Javascript code and other client-side techniques that are not amenable to web harvesting. Because of this many web harvesters do a poor job of capturing social media websites (Thomson 2016).

● Social media data can generally be collected more efficiently from the API than the website. Some APIs expose API methods that crosscut the social media data in ways

3 http://json.org/
4 https://github.com/internetarchive/heritrix3
5 https://webrecorder.io/
6 https://github.com/docnow/twarc. Twarc is a tool for collecting from the Twitter API. While web harvesting software is generally intended to collect content from any website, social media harvesting software is not general-purpose; it can usually only collect from one to a small number of APIs.
that aren’t possible from the website. For example, via its API, Twitter allows filtering the stream of all tweets currently passing through the platform.

Some of the weaknesses of collecting from an API include:

- Not all social media platforms have complete, public APIs. This will be discussed further below as it relates to the Weibo API.
- Each API is different.
- Some platforms, notably Twitter, place explicit requirements and limits on the use and sharing of data harvested using the API. These may be more explicit or limiting than the terms of service for the websites.
- Limitations placed by the platform on the amount of data that can harvested via the API can make it difficult or impossible to capture older content.

Affordances of the Weibo API
With this basic understanding of collecting from the API of a social media platform, let’s consider the affordances of the Weibo API. For starters, here is part of a weibo retrieved from the Weibo API:

```json
{
  "created_at": "Thu May 11 15:43:13 +0800 2017",
  "id": 4106245138832598,
  "text": "民生服务
【中央机关公开遴选选调360名公务员】2017年中央机关公开遴选和公开选调公务员工作今日开始报名。此次公开遴选和公开选调共有56个中央机关参加计划选拔360名公务员。[心]报名时间截止5月22日18:00。[心]笔试时间为2017年6月25日[心]考试地点设在北京、上海、西安、兰州等17个城市。详情",
  "textLength": 314,
  "source": "<a href="http://app.weibo.com/t/feed/6ghA0p" rel="nofollow">搜狗高速浏览器</a>",
  "favorited": false,
  "truncated": false,
  "in_reply_to_status_id": "",
  "in_reply_to_user_id": "",
  "in_reply_to_screen_name": "",
  "geo": null,
  "user": {
    "id": 5083132536,
    "screen_name": "西安国际港务区",
```

7 The full JSON for this weibo is available at: https://gist.github.com/justinlittman/db0280402a2ded54cc3d539a798c16e0. For comparison, here is the JSON for a tweet: https://gist.github.com/justinlittman/462a398d161002a8caff0905bf4e5f7f.
The meanings of some fields are described in the API documentation8 (Sina Corp 2017).

Here is how this weibo is rendered on the Sina Weibo website:

---

8 http://open.weibo.com/wiki/2/statuses/home_timeline
As mentioned previously, the weibo is in JSON, a simple format that is widely used to exchange data on the web. In the case of a weibo, some of the more significant fields include:

- **created_at**: A timestamp for when the weibo was posted.
- **id**: A unique identifier for the weibo.
- **text**: The text of the weibo.
- **user**: Information on the author.
- **reposts_count**: The number of times the weibo has been reposted by other users. This is as of the time that the weibo was retrieved from the API and may change over time.
- **comments_count**: The number of comments by other users on the weibo. This also may change over time.
- **geo**: Optional geotagging of the location from which the weibo was posted.
- **isLongText**: Marks whether the JSON text field is truncated text or is long text. When true, the text of the weibo contained in the text field is truncated to 140 characters. The API does not provide a mechanism for getting the complete text of the weibo, which is a significant limitation.

Note that in this example, the text of the weibo is 161 characters, while the entire JSON-formatted weibo is 10,460 characters.
Like other social media APIs, Sina Weibo provides a number of different methods for interacting with the API. These include functions like “statuses update” for posting a weibo and “friendships destroy” for unfollowing another user (again, think the functions that would be needed by a Weibo smartphone app). For the purposes of collecting social media data, our focus is on methods that allow retrieving weibos. Unfortunately, this is one of the fundamental challenges of collecting content from the Sina Weibo API: the methods for retrieving weibos are extremely limited.

The Weibo API is split into basic methods and advanced methods. Anyone can get access to basic methods; access to the advanced API involves more hurdles and will be discussed further below.

The basic API provides the “friends timeline” method, which returns the weibos of the user and the user’s friends. (On Sina Weibo, users can have followers. These are more like following a user on Twitter than having a friend on Facebook, as the assent of the friend is not necessary to establish the friend relationship.) The “friends timeline” method can be used as indirect mechanism for collecting the posts of a set of users. So, for example, for the anti-corruption campaign we want to collect the posts of a set of users that we identified that post about the topic. To use the friends timeline method to collect weibos for a single collection, a new, “dummy” account must be created and that account must follow only the desired users.

Using the “friends timeline” method to collect the tweets of other users can be contrasted with Twitter API’s “statuses/user_timeline” method. The “statuses/user_timeline” method allows collecting the most recent tweets of any user (if the account is not protected). It does not require the indirection of having to have a “dummy” account follow the users. Put more directly, this approach to collecting with the “friends timeline” method is a creative “kludge” due to the limitations of the Sina Weibo API.

Based on experimentation, we have determined that the friends timeline method only returns the last 150 weibos from the user and the user’s friends. (Sometimes it might be 1 or 2 less due to unknown reasons). For example, if User 张三 posts 200 weibos and User 李四

---

9 For a complete list, see http://open.weibo.com/wiki/%E5%BE%AE%E5%8D%9AAPI.
10 In the documentation, advanced methods are indicated by “高” (advanced).
11 http://open.weibo.com/wiki/2/statuses/friends_timeline
12 It might seem like the “user timeline” method (http://open.weibo.com/wiki/2/statuses/user_timeline) would be useful for collecting. However, this API call only return the weibos of the authorized user, requiring that each users need to authorize your application to collect his timeline. This isn’t feasible for large scale collecting.
posts 150 weibos, the result would be the 150 latest weibos among the total 350 weibos. To attempt to compensate for this, weibos can be collected from the “friends timeline” on a regular schedule, e.g., every half hour. Still, weibos collected via this method should be considered a sample and not a complete set of all weibos authored by the user and the user’s friends.

The advanced API provides the “topic search” method\footnote{http://open.weibo.com/wiki/2/search/topics}, which returns the most recent 200 posts matching a topic (again, sometimes minus 1 or 2 posts due to unknown reasons). A topic is a single keyword bracketed by “#” in the text of a weibo, e.g., “#keyword#” or “#你好#”; topics on Weibo are like hashtags on Twitter, although they are bracketed by hashtag symbols rather than preceded by a single hashtag symbol.\footnote{The closing hashtag is needed due to the lack of spacing in Chinese characters.} Marking keywords as topics allows users to contribute to a folksonomy on the Sina Weibo platform, as other users can discover their weibos based on the presence of the topic. Since each call to the "topic search" API only returns a limited number of weibos, collecting on a regular schedule is essential to building as complete a topic collection as possible.

Compared to a social media platform such as Twitter, Sina Weibo is rather limited in terms of the API methods that would be desired in order to perform comprehensive collecting. Still, as will be shown below, given the volume of weibos passing through the Sina Weibo platform and when coupled with software such as SFM, the Weibo API does permit creating large scale and valuable datasets for research.

**Social Feed Manager**

Social Feed Manager\footnote{For project information see http://go.gwu.edu/sfm. For software documentation, see http://sfm.readthedocs.io/en/latest/. For a conceptual and technical overview of SFM, see (Littman et al. 2016).} (SFM) is an open source software developed by GW Libraries\footnote{SFM is supported by grant #NARDI-14-50017-14 from the National Historical Publications and Records Commission.} that harvests social media data and web resources from Twitter, Tumblr, and Flickr, in addition to Sina Weibo.\footnote{The Sina Weibo harvester (https://github.com/gwu-libraries/sfm-weibo-harvester) was developed by Yecheng Tan.} SFM empowers researchers, faculty, students, and archivists to define and create collections of social media data. While SFM can be run on a laptop, it is intended to be hosted on a server and operated by a collecting organization as a service for the members of its community. For example, GW Libraries operates an instance of SFM that is available to campus researchers, including students, faculty, and library staff.
One of the primary goals of SFM is to lower the barriers to collecting social media data. This has two aspects: First, users interact with a user-friendly website to create, manage, and export their collections. Users do not need to have technical skills or detailed knowledge of social media APIs (although an understanding of the affordances of the APIs is recommended for successful research). Extensive documentation\textsuperscript{19} has been created to assist users in creating social media collections with SFM (GW Libraries 2015).

Second, deploying SFM has been made as simple as possible. Using a technology called Docker, SFM can easily be deployed on a server with a minimal number of steps. (Instructions are also provided for deploying “in the cloud” using Amazon Web Services.) Still, SFM does require a moderately technical administrator and the availability of a server. Just as for users, extensive documentation is provided for administering SFM\textsuperscript{20}.

Setting up a Sina Weibo collection in SFM involves five steps. The first step is to create an account on Sina Weibo’s website and follow the users that you would like to collect. The next step is to provide a credential (aka, API key). Like other social media platforms, Sina Weibo requires a credential to access the API. (The challenges of acquiring a credential will be discussed below.) Once a user has a credential, she supplies it to SFM so SFM can use it for collecting.

\textsuperscript{19} http://sfm.readthedocs.io/en/latest/#user-docs
\textsuperscript{20} http://sfm.readthedocs.io/en/latest/#admin-documentation
The third step is to create a collection set. A collection set is merely a container for collections. In the case of China's anti-corruption campaign, we have a collection set that contains a Sina Weibo timeline collection, two Twitter collections, and a number of Sina Weibo topic searches. (The list of collections is truncated in this screenshot.)
The fourth step is to create a collection. Each collection is for a particular part of a single social media API. For example, for Twitter, SFM supports four different collection types (user
timeline, search, filter, and sample). For Sina Weibo, SFM supports timeline and search\textsuperscript{21} collection types corresponding to the API methods described previously.

For each collection, the user selects the credential to be used, a schedule (e.g., every 30 minutes, every day, every week), and the collecting parameters (e.g., whether to collect web resources for images or URLs found in weibos).

\textsuperscript{21} For reasons that will be obvious when we discuss Sina Weibo challenges, the search collection type is not enabled by default.
Most other types of collections require the user to provide seeds. For example, a Twitter user timeline collection requires the screen names of the accounts to collect (e.g., @SocialFeedMgr). However, since the weibos that are collected for a weibo timeline collection are determined by the friends of the user whose credentials are used, seeds are not applicable.

Lastly, the collection must be turned on.
SFM initiates harvests as determined by the schedule. In these harvests, SFM's "harvester" components request content from the social media APIs. SFM provides the user with information about the harvests that were performed.

<table>
<thead>
<tr>
<th>Type</th>
<th>Requested</th>
<th>Updated/Completed</th>
<th>Status</th>
<th>Stats</th>
<th>Messages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 10:23:51 a.m. EDT</td>
<td>May 12, 2017, 10:24:00 a.m. EDT</td>
<td>Success</td>
<td>24 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 9:23:51 a.m. EDT</td>
<td>May 12, 2017, 9:24:00 a.m. EDT</td>
<td>Success</td>
<td>27 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 8:23:51 a.m. EDT</td>
<td>May 12, 2017, 8:24:00 a.m. EDT</td>
<td>Success</td>
<td>23 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 7:23:51 a.m. EDT</td>
<td>May 12, 2017, 7:24:00 a.m. EDT</td>
<td>Success</td>
<td>25 weibos</td>
<td></td>
</tr>
<tr>
<td>Weibo timeline</td>
<td>May 12, 2017, 6:23:51 a.m. EDT</td>
<td>May 12, 2017, 6:24:00 a.m. EDT</td>
<td>Success</td>
<td>51 weibos</td>
<td></td>
</tr>
</tbody>
</table>

View all 20,980 harvests

SFM also provides summary statistics for a collection.

**Description:** Followers List: 反腐败斗争的博客 法治说 刘荣香举报省农垦局领导腐败 反腐败 反贪局长张秀芳 打假反腐专员 孟治 腐败引发民愤 涨腐败 箭门腐败 全民反腐腐败 史海觅珠 开封检察 中纪委关注农村腐败 南方都市报 人民日报 人民网 新闻调查卧底反腐记者张子保达人 反腐倡廉网络舆情 红色情感 腐败反腐联合指挥部 新闻（New Add:2017-01-30） 南方周末 法制晚报 央视新闻 澎湃新闻 国际反贪 法制网 最高人民检察院 南方日报 新浪财经 检察日报 人民网 丰都检察

**Data collected:** 19,376 files (26.0 GB)

**Stats:**
- weibos: 92,497
- web resources: 306,032

There are several mechanisms for accessing social media data. By far the most common and simplest is to export it to a spreadsheet format such as Excel or CSV. Other options include a JSON-formatted export, or access from the command line.
Key fields are extracted from each weibo to flatten it into a spreadsheet row. For weibos, this is documented in the data dictionary.22

Challenges in working with the Weibo API

Getting SFM to this point in supporting collecting social media data from Sina Weibo required navigating a number of challenges. The first challenge is the Weibo documentation, which is available in Mandarin23 and English24 (Sina Corp 2017). The English documentation is sorely out of date. While the Mandarin documentation is less out of date, it seems to often be inaccurate. The SFM team fortunately includes a native Chinese speaker and was able to overcome the errors in the documentation through careful experimentation.

The second challenge is the process of getting credentials, especially for those not fluent in Mandarin. While we did encounter some existing English instructions for creating an account and applying for credentials, they were out of date. To support SFM users and other Sina Weibo researchers, we wrote our own instructions (Tan 2016).

Applying for credentials to the advanced API requires some additional steps (documented in the instructions mentioned above). The software application that is requesting the

---

23 http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5
24 http://open.weibo.com/wiki/%E9%A6%96%E9%A1%B5/en
credentials must have at least 1,000 authorized users. This is a considerable hurdle; options include manually creating accounts or purchasing accounts from the grey market that exists around Sina Weibo and then performing authorization.\textsuperscript{25} There is no easy way to automate authorization due to the anti-robot/CAPTCHA. It should be recognized that this represents a significant barrier to other institutions using the topic search API method to collect weibos.

And lastly, for security reasons some parts of the Sina Weibo website are blocked by some organizations. (These are primarily resources included in web pages, not the site itself.) In the case of GW, we could not register for credentials on the GW campus network; we had to use an external network.

**China’s anti-corruption campaign collection overview**

While the social media data related to the Chinese anti-corruption campaign has not yet been closely studied, we can at least consider an overview of the collection. Since its inception on June 29, 2016 until May 10, 2017, 91,584 weibos and 23,218 tweets have been collected.

This visualization show the monthly collection rate over the past year.

Here are the top 10 topics from the past year in the dataset:

\textsuperscript{25} In early 2017, purchasing accounts used for authorization on Taobao (https://www.taobao.com) cost around $1 per 100 accounts.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>纪检监察</td>
<td>545</td>
</tr>
<tr>
<td>反腐倡廉</td>
<td>457</td>
</tr>
<tr>
<td>2017两会</td>
<td>417</td>
</tr>
<tr>
<td>2017看两会</td>
<td>402</td>
</tr>
<tr>
<td>北电学生举报教授</td>
<td>401</td>
</tr>
<tr>
<td>央视新闻微直播</td>
<td>380</td>
</tr>
<tr>
<td>关注里约奥运</td>
<td>373</td>
</tr>
<tr>
<td>微博看两会</td>
<td>294</td>
</tr>
<tr>
<td>北电学生实名举报女教授</td>
<td>264</td>
</tr>
<tr>
<td>一起看奥运</td>
<td>245</td>
</tr>
</tbody>
</table>

When we take a deep look at the top 10 topics, 纪检监察, 反腐倡廉, 北电学生举报教授 and 北电学生实名举报女教授 are closely related to the anti-corruption campaign. The other three topics, 2017两会2017看两会 and 微博看两会, are all related to 两会26 which refer to two annual political plenary sessions, National People’s Congress and National Committee of the Chinese People’s Political Consultative Conference. In both conferences, the anti-corruption campaign is one of the most important topics. 央视新闻微直播 is the official Sina Weibo account of CCTV’s center for news which covers breaking news and featured news. News related to the anti-corruption campaign is probably one of its foci.

26 https://en.wikipedia.org/wiki/Lianghui
And a word cloud of the top 30 topics:

Over the past year, the most prolific weibo authors have been:

<table>
<thead>
<tr>
<th>Screen name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>人民网</td>
<td>21354</td>
</tr>
<tr>
<td>人民日报</td>
<td>11925</td>
</tr>
<tr>
<td>刘茉香举报省农垦局领导腐败</td>
<td>7817</td>
</tr>
<tr>
<td>打腐败促正义</td>
<td>7805</td>
</tr>
<tr>
<td>新浪财经</td>
<td>6804</td>
</tr>
<tr>
<td>新闻调查反腐卧底记者张子保</td>
<td>6404</td>
</tr>
</tbody>
</table>
Among the 10 accounts, 人民网人民日报央视新闻 represent official government media, which reports comprehensive news that covers the anti-corruption campaign. 新浪财经 is Sino’s official account with a focus on finance and economics, and also covers news on corruptions. 法制网 and 法制晚报, as the most prestigious media on law, are major media channels revealing corruption cases. 刘茉香举报省农垦局领导腐败, 打腐败促正义 and 新闻调查反腐卧底记者张子保 are private accounts that focus exclusively on anti-corruption.

The top 10 links includes in weibos were:

<table>
<thead>
<tr>
<th>Link</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://m.weibo.cn/status/3987396334993894?wm=3333_1001">https://m.weibo.cn/status/3987396334993894?wm=3333_1001</a></td>
<td>573</td>
</tr>
<tr>
<td><a href="http://rp11047911.faisco.cn/">http://rp11047911.faisco.cn/</a></td>
<td>491</td>
</tr>
<tr>
<td><a href="https://m.weibo.cn/status/3983754928012045?luicode=2000061&amp;lfid=3987396334993894&amp;featurecode=20000180&amp;wm=2468_1006">https://m.weibo.cn/status/3983754928012045?luicode=2000061&amp;lfid=3987396334993894&amp;featurecode=20000180&amp;wm=2468_1006</a></td>
<td>388</td>
</tr>
<tr>
<td><a href="https://m.weibo.cn/status/3987396334993894?luicode=2000061&amp;lfid=3983754928012045&amp;featurecode=20000180&amp;wm=3333_1001">https://m.weibo.cn/status/3987396334993894?luicode=2000061&amp;lfid=3983754928012045&amp;featurecode=20000180&amp;wm=3333_1001</a></td>
<td>383</td>
</tr>
<tr>
<td><a href="http://weibo.com/ttarticle/p/show?id=2309403980091532139574">http://weibo.com/ttarticle/p/show?id=2309403980091532139574</a></td>
<td>317</td>
</tr>
<tr>
<td><a href="https://m.weibo.cn/status/4044157926040825?luicode=2000061&amp;lfid=3987396334993894&amp;featurecode=20000180&amp;wm=2468_1007">https://m.weibo.cn/status/4044157926040825?luicode=2000061&amp;lfid=3987396334993894&amp;featurecode=20000180&amp;wm=2468_1007</a></td>
<td>270</td>
</tr>
</tbody>
</table>
Note that some of these links may require a Sina Weibo account; others may no longer be available.

And lastly, these are the top 10 users mentioned over the past year:

<table>
<thead>
<tr>
<th>User</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>明月清风74545</td>
<td>1584</td>
</tr>
<tr>
<td>打腐败促正义</td>
<td>1347</td>
</tr>
<tr>
<td>中国政府网</td>
<td>986</td>
</tr>
<tr>
<td>感恩</td>
<td>969</td>
</tr>
<tr>
<td>素素abc</td>
<td>676</td>
</tr>
<tr>
<td>少佳小屋2</td>
<td>520</td>
</tr>
<tr>
<td>中国政府网</td>
<td>518</td>
</tr>
<tr>
<td>最高人民检察院</td>
<td>496</td>
</tr>
</tbody>
</table>
Four of the top five users are private accounts (including 明月清风74545 打腐败促正义 感恩 素素 abc) that focus is on anti-corruption. They follow each other, and repost each other's weibos. 中国政府网 最高人民检察院 and 人民网 are official accounts of the Chinese government.

Though only a cursory look at the anti-corruption campaign dataset, it indicates that collecting has largely been on-topic. Further, it hopefully demonstrates the research potential of the dataset.

Conclusion
When the social media data is combined with the blogs captured by our colleagues, we contend that this represents a unique and valuable collection for current and future scholars to study non-official voices in China’s anti-corruption campaign. Though there are significant limitations to the Sina Weibo API and challenges in using the API, we believe that our work in adding Sina Weibo support to SFM and building the China’s anti-corruption campaign collection proves that these can be overcome to perform meaningful collection building. Further, we are hopeful that through the demonstration of this collection and by lowering the technical barriers to collecting from Sina Weibo, it will be possible for others in the community to draw from the millions of Sina Weibo users to create other compelling research collections.

Acknowledgements
The authors are grateful for the support of the Mellon Foundation-CEAL Innovation Grant Program and the opportunity to collaborate with Yunshan Ye and Ding Ye. In addition, we appreciate the contribution of current and former colleagues on the SFM team: Laura Wrubel, Aditya Dharne, Christie Peterson, Dan Chudnov, Rachel Trent, Soomin Park, and Yonah Bromberg Gaber.
References


NEW APPOINTMENTS

Tang Li joined the East Asian Library at the University of Southern California as Chinese Studies Librarian as of July, 2017. She joins a team of four other librarians and will be responsible for collection management for Chinese materials in all formats, as well as for reference and liaison services between the library and the faculty and students involved in Chinese Studies at USC. She worked at Yale University’s East Asia Library for the previous nine years and is currently Chair of CEAL’s Library Technology Committee. Tang can be contacted by email at tangl@usc.edu or by phone at (213) 740-8548.

(From an Eastlib posting by Ken Klein, University of Southern California)

Stella Tang has been appointed Chinese Cataloging Librarian at the C. V. Starr East Asian Library at UC Berkeley for a term of three years. Stella received her MLS from the University of Michigan in 1980. Since then, she has held professional positions in technical services in Washington, DC, and in the Bay Area, chiefly at SRI in Menlo Park, the Olin Library at Mills College, and the California Academy of Sciences. She is a member of ALA, CEAL, and a number of specialized professional organizations. Ms. Tang can be reached at 510-642-9220 or by email at stellat@berkeley.edu.

(From an Eastlib posting by Haiqing Lin, University of California, Berkeley)

The 50th anniversary of the Chinese Cultural Revolution in 2016 inspired several commemorative seminars and conferences in Germany, France, Japan, Hong Kong and the United States. However, only one such commemorative event held in Los Angeles and UC Riverside entitled *China and Mao's Legacy: Commemoration of the 50th Anniversary of the Cultural Revolution* produced a two-part conference volume entitled *China and the Maoist Legacy: The 50th Anniversary of the Cultural Revolution*. Complete with nearly one million Chinese characters and edited by Yongyi Song, a leading Cultural Revolution scholar and one of the chief organizers of the conference, this conference volume is a significant contribution to the existing literature on the Cultural Revolution.

Since such scholarly activities and exchanges are restricted in China, conference organizers felt duty-bound to convene such a conference overseas. Their efforts came to great fruition. This volume contains twenty essays from a great many prominent Cultural Revolution scholars and researchers from mainland China, including Xu Youyu, Qin Hui, Yang Jisheng, Wang Haiguang, Yin Hongbiao, Mi Hedou, Zhu Xueqin, Tang Shaojie, Dong Guoqiang, Shen Xiaoyun, Xu Hailiang, Jin Guangyao and Li Xun. In addition, over fifty scholars from the United States, Canada, Germany, Japan, Hong Kong, Macau, and other parts of the world also gathered at the conference and contributed their research papers to this volume, providing a well-rounded representation of recent and current research findings on the Cultural Revolution.

Another important aspect of this conference volume is that the views of some emerging scholars on the Cultural Revolution are heard and represented. Over the last decade, many young scholars and doctoral students who have focused on Cultural Revolution studies have emerged both inside and outside China. Recognizing the importance of this community, the editor has included a few essays from the so-called “ba ling hou” generation (people born in the decade of 1980s). Six nascent researchers from various parts of the world contributed papers on such Cultural Revolution-related topics as work teams, violence, the handling of grievances in Guangxi, and a comparison of worker rebels in Shanghai and Wuhan. Although these young scholars are all members of the post-Cultural Revolution generation, their research has nonetheless reflected solid command of archival sources and knowledge of new scholarship. Their contribution is pivotal for the future of Cultural Revolution studies.

Yet another important contribution of this volume is the representation of new historical facts, perspectives, and theories, as reflected in the writings of a number of prominent scholars, including Xu Youyu, Song Yongyi, Wu Guoguang, Cheng Yinghong, Yang Haiying, Ding Shu, Mi Hedou, Hu Pin, Cheng Tijie, and Zhou Zehao. Efforts to establish new
Some historical events during the Cultural Revolution are familiar to many, such as the “February Suppression of Counterrevolutionaries,” “Extensive Democracy,” and “Elections at Party Congresses.” However, in-depth research on them has long been absent. Professor Xu Youyu’s paper “February Suppression of Counterrevolutionaries” was the first effort to comprehensively investigate and shed light on this major historical event through a combination of broad analysis and careful documentation. For another instance, Professor Wu Guoguang of the University of Victoria delved into the phenomenon of “Extensive Democracy” with which the Cultural Revolution is often associated. In his paper “Elections during the Cultural Revolution: An Investigation of Party Congress,” Wu exposed the nature of “Extensive Democracy” during the Cultural Revolution through meticulous analysis.

Notwithstanding the restrictions on Cultural Revolution studies, the legacy of the Cultural Revolution continues to be keenly felt in contemporary China. It is inevitable, therefore, that any meaningful research on the Cultural Revolution and its political legacy must include an assessment of the impact of the Cultural Revolution on China today. Several papers in this volume address this topic through in-depth discussions. For example, Gao Wenqian’s paper studied the parallels between some practices during Cultural Revolution and the leadership style in contemporary China while Ling Cangzhou explored the impact of Mao’s legacy on contemporary China.

The editor also created a special section titled “The Reexamination of the Cultural Revolution Studies” to sum up the state of research on this historical event. In this section, Wang Haiguang, a professor from the History Department of the East China Normal University, contributed his paper “Cultural Revolution Research Yesterday, Today, and Tomorrow: a Comprehensive Review of Research on the Cultural Revolution,” in which he described and reflected on the long and difficult trajectory of Cultural Revolution research in mainland China. Professor Guo Jian of the University of Wisconsin-Whitewater presented a paper titled “Rebel Complex.” Guo argued that the healthy
development of a research field is improbable without frequent introspection and self-corrections on the part of the participants of the historical events in question. All these perspectives have led the discussions into many hitherto unexplored areas and served to both enrich the content and broaden the scope of this volume.

In sum, the publication of this conference volume will not only promote Cultural Revolution studies as an academic field but it will also provide much needed clarity and perspectives on Mao's legacy and its continuous impact on China today and tomorrow. This conference volume will be a welcome addition to any Cultural Revolution collection in academic libraries as well as to any public library with a comprehensive Chinese collection.

Heng Ge, NYU Shanghai Library