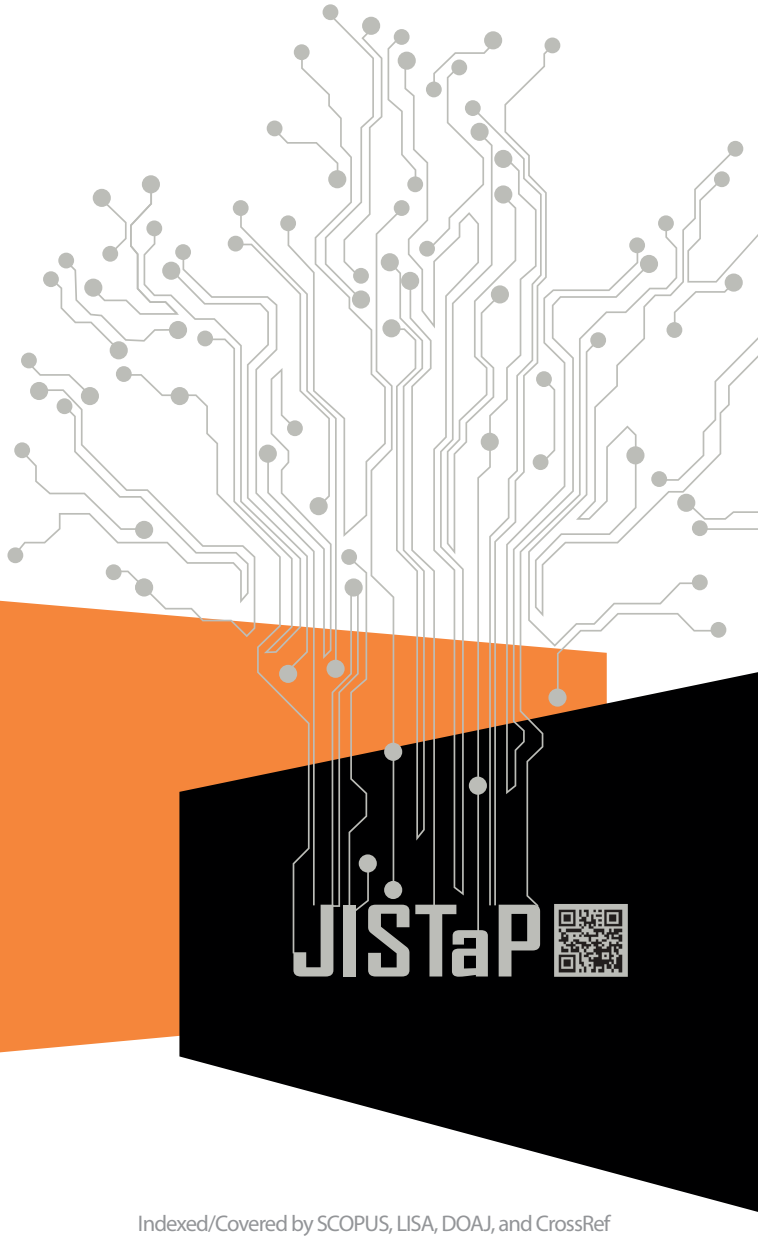
eISSN : 2287-4577 pISSN : 2287-9099
http://www.jistap.org
Vol. 7 No. 1 March 30, 2019

Journal of Information Science Theory and Practice



JISTaP

KISTI www.kisti.re.kr

Indexed/Covered by SCOPUS, LISA, DOAJ, and CrossRef

General Information

Aims and Scope

The *Journal of Information Science Theory and Practice (JISTaP)* is an international journal that aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics. JISTaP will be published quarterly, issued on the 30th of March, June, September, and December. JISTaP is indexed in the Scopus, Korea Science Citation Index (KSCI) and KoreaScience by the Korea Institute of Science and Technology Information (KISTI) as well as CrossRef. The full text of this journal is available on the website at http://www.jistap.org

Indexed/Covered by

Publisher

Korea Institute of Science and Technology Information
66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea
(T) +82-2-3299-6102
(F) +82-2-3299-6067
E-mail: jistap@kisti.re.kr
URL: http://www.jistap.org

Managing Editor: Ji-Young Kim, Eungi Kim

Copy Editor: Ken Eckert

Design & Printing Company: SEUNGLIM D&C

4F, 15, Mareunnae-ro, Jung-gu, Seoul, Republic of Korea
(T) +82-2-2271-2581~2
(F) +82-2-2268-2927
E-mail: sdnc@sdnc.co.kr

Open Access and Creative Commons License Statement

This paper meets the requirements of KS X ISO 9706, ISO 9706-1994 and ANSI/NISO Z39.48-1992 (Permanence of Paper)

Table of Contents

JISTaP

Vol. 7 No. 1 March 30, 2019

Journal of Information Science Theory and Practice • http://www.jistap.org

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 06-16, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.1

Journal Publishing and Authorship in Library and Information Science by Early Career Researchers in South Korea

Eun-Ja Shin*

Department of Media and Communication, Sejong University,
Seoul, Korea
E-mail: ejshin@sejong.ac.kr

ABSTRACT

This study explored journal publishing and authorship by South Korean early career researchers (ECRs) in the field of library and information science (LIS). This research analyzed relevant journal publication data and conducted interviews to obtain information on the experiences and opinions of ECRs. Results indicated that South Korean ECRs in LIS were highly productive. This was evidenced by their annual publishing rate of 2.04 articles per person. In addition, Social Science Citation Index (SSCI) publications were produced at an annual average of 0.26 articles per person, while the quartile ratings for SSCI journal publications were also relatively high. However, unlike the trends seen in other academic fields, their collaborative research efforts were not considered very high because such efforts did not correspond to half their total publications. ECRs often participate as lead or corresponding authors despite being new researchers. ECRs are publishing first in the journals approved by their universities. These researchers cannot receive proper credit if the journal was not approved in this manner. ECRs are particularly disadvantaged when publishing in international journals corresponding to specific areas that are not on the SSCI list. By examining the journal publishing and authorship efforts of ECRs, this study discovered a variety of difficulties that should be addressed. For example, South Korean universities do not currently have cooperative research guidelines to solve authorship problems. The results from this study can serve as a basis to establish academic publishing and authorship policies while promoting scholarly communication in LIS and other scientific fields.

Keywords: early career researcher, journal publishing, authorship, Social Science Citation Index, library and information science

Open Access

Accepted date: January 27, 2019
Received date: September 01, 2018

***Corresponding Author:** Eun-Ja Shin
Full professor
Department of Media and Communication, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Korea
E-mail: ejshin@sejong.ac.kr

1. INTRODUCTION

Research competition has intensified throughout the global era, while academic publishing has also shown significant change. In particular, there is now a greater quantity and quality of academic publications from new researchers (Shaw & Vaughan, 2008; Lee & You, 2014; Choi & Yang, 2018). This study aimed to determine the current conditions of journal publishing and authorship for early career researchers (ECRs) while identifying any related problems.

The scholarly definition for ECR differs from that used in this study. For example, Nicholas et al. (2017) and Xu, Nicholas, Zeng, Su, and Watkinson (2018) defined ECRs as researchers not exceeding 35 years of age who have either earned a doctoral degree or are currently enrolled in a PhD program. Both studies explained that ECR status corresponds to instability for researchers who have not achieved full-time or tenured employment. However, this study defined ECR in accordance with the application requirements for the young researcher program at the National Research Foundation of Korea (NRF, https://www.nrf.re.kr/; i.e., researchers who earned a doctoral degree less than 10 years prior to applying or were employed in colleges or universities as assistant professors for less than five years). This study did not establish an age limit and extended the definition of ECR to include postdoctoral fellows and beginning professors. Regardless of whether they are unemployed or already hired, ECRs are not in a secure position to consistently publish papers for contracting and promotion.

Assistant professors working in South Korea spend a significant amount of time in education, research, and service delivery for promotion evaluations. ECRs may have trouble in the classroom because they have little teaching experience and sometimes face obstacles in communicating with students. Negative end-of-semester lecture evaluations may negatively affect contracting and promotion. ECRs especially focus on teaching for this reason. In fact, it is journal publishing that significantly impacts recruitment, contracting, promotion, and tenure consideration.

South Korean universities and research institutes have emphasized the quantitative aspects of research achievements for decades. However, qualitative aspects have also gained consideration in recent years. Academic institutions largely believe that it is reasonable to gauge the quality of journals according to citation counts. Researchers at academic institutions are thus pressured to publish in high-ranking journals with high citation counts. High-ranking journals often refer to top journals indexed in the Web of Science (WoS, http://www.webofknowledge.com) or Science Citation Index (SCI)/Social Science Citation Index (SSCI)/Arts & Humanities Citation Index (A&HCI). The journal impact factor quartile for publications provided by Journal Citation Reports (JCR, https://jcr.incites.thomsonreuters.com) does not differ from the calculation method. South Korean academic institutes generally prefer to publish in first or second quartile journals. There is also concern that publishing in fourth quartile journals will result in negative consequences for the university's rating, especially in the Leiden ranking.

The 2018 Leiden ranking includes 938 global universities that have produced at least 1,000 WoS indexed publications in the last four years (Leiden Ranking, 2018). In addition, the number of publications in the top 10% of total papers is an important criterion. Thus, having more papers in the top 10% results in a better Leiden ranking. Ironically, not publishing is more advantageous to achieving a positive Leiden assessment than publishing papers that receive few citations. South Korean universities have recently begun to promote publication in international journals while producing highly-cited papers and conducting collaborative research. ECRs are therefore highly motivated and aware of the need to publish highly cited papers in high-ranking journals.

This atmosphere has urged South Korean ECRs to constantly publish in international journals. This is unlike the situation for senior researchers, who mainly target South Korean journals (Lee & Bak, 2016). It is not easy for ECRs to conduct research; they do not have the requisite experience. Many also have difficulty publishing in English as opposed to their native language. University commitments to publishing (especially in high-ranking journals) are thus likely to put pressure on ECRs. Meanwhile, scholarly communication has become much more active because of various widespread media practices, including those seen through social networking services and websites like YouTube (Brand, Allen, Altman, Hlava, & Scott, 2015). The proportion of collaborative research has also increased. These are the current trends affecting the academic environment for ECRs. South Korean social norms tend to ensure that main positions and roles are given to those with more seniority (i.e., "age before honesty"), which may also affect author role distributions in collaborative papers. Senior researchers may have advantages when assigning main authors, while ECRs can feel that they do not receive similar chances (Maciejovsky, Budescu, & Ariely, 2009). This academic situation and the associated social norms provide a background for studying ECR journal publishing and authorship.

Journal publishing patterns and authorship practices vary widely across disciplines. It is thus undesirable to analyze the situation in a general sense. A more appropriate analysis

involves data examination to derive implications according to specific disciplines. This study, therefore, elected to analyze ECRs from the field of library and information science (LIS). Theories are also important in this area. However, it is highly necessary to gain feedback from librarians in this field regarding their experiences in achieving academic advancement. Librarians, policymakers, and professors working in LIS often collaborate through industrial-academic projects. Here, the influence of project managers can become significant, resulting in disadvantages for ECRs or librarians in terms of authorship. It is also the responsibility of scholars in the scientific community to establish a reasonable authorship-credit allocation policy and refine cooperative research guidelines for addressing the authorship problem (Brand et al., 2015).

In this context, this study reviewed previous research on journal publishing in the LIS field before collecting and comprehensively analyzing ECR journal publishing and authorship data. Interviews were then conducted with ECRs. The results will serve as a basis for establishing academic publishing and authorship policies in addition to promoting scholarly communication in LIS and other scientific fields.

2. LITERATURE REVIEW

Few previous studies have focused on publishing productivity in the LIS field, especially regarding the activity of ECRs. This study, therefore, broadened the scope by examining research productivity and authorship in LIS regardless of researcher age or position.

Many studies use paper counts to measure the research productivity of LIS authors (Adkins & Budd, 2006; Choi & Yang, 2018; Chung & Park, 2011; Davarpanah & Aslekia, 2008; Larivière, Sugimoto, & Cronin, 2012; Lee & Yang, 2011a; Lee & Yang, 2011b; Shaw & Vaughan, 2008). This study relied on Lee and Yang's (2011b) research on the journal publishing productivity of South Korean LIS professors, in which they examined papers published by 159 South Korean LIS professors from 2001 to 2010. These professors published 2,231 papers in national journals and 111 papers in international journals. The annual average number of publications per person was 1.40 in South Korean journals and 0.07 in international journals. Of those studied, 36 professors were published in WoS journals (22.64% of the total).

Choi and Yang (2018) showed the number of papers produced by 205 South Korean LIS professors from 2011 to 2016. Of these, 1,789 papers were published in national journals, while 221 were published in international journals. The annual average number of publications per person was 1.45 in South Korean journals and 0.18 in international journals. Both of the above studies indicated that while the number of publications produced by South Korean LIS professors in South Korean journals remained nearly unchanged, their number of publications in international journals increased significantly.

Shaw and Vaughan (2008) analyzed the journal publishing of 720 LIS professors at universities in the United States during their active scholarship lifetime. They revealed that the average annual number of publications in print journals per person was 0.25 for assistant professors, 0.35 for associate professors, and 0.72 for full-time professors. Adkins and Budd (2006) examined SSCI articles written by LIS professors in the United States from 1999 to 2004. The most productive professors produced a very large number of papers (e.g., Tenopir with 59, Jasco with 32, and Cronin with 25). Meanwhile, Mukherjee (2010) examined LIS SSCI papers from Asian authors who published between 2001 and 2007. He also included LIS articles written by researchers within other majors. All totalled, 384 were written in China, 275 in Taiwan, and 216 in South Korea. The paper counts in all three countries significantly increased from 2001 to 2007.

Lee and Yang (2011a) presented the number of joint studies performed by 159 South Korean LIS professors. From 2001 to 2010, 52.75% of all papers were produced by a single author, while 30.75% were produced by two authors (i.e., 83.5% of these publications were written by one or two authors). Choi and Yang (2018) counted publications by 205 South Korean LIS professors between 2011 and 2016 and found that although the number of single-authored papers decreased, the number of collaborative papers increased. Their charts show that the number of single-authored papers decreased from about 56% in 2011 to approximately 38% in 2016. This indicates that instances of collaborative research have significantly increased. Of note, the number of collaborative papers consisting of more than three authors exceeded 10% of all papers in 2015 and 2016. The popularity of collaborative research is increasing not only in the LIS field but also in other fields around the world. Frandsen and Nicolaisen (2010) conducted an analysis of joint research papers published in 12 information science core journals from 1978 to 2007. Results showed that multi-authorship had increased over time. Their charts showed that in 1978, approximately 64% of all studied papers were from single authors, 27% were from two authors, and 8% were from three authors. In 2007, approximately 33% of all studied papers were from a single author, 31% were from two authors, and 23% were from three authors.

At any university, a single author can obtain a 1 (i.e., 100%

with a total credit). However, the credit allocation method used for co-authors varies greatly depending on the university. Lee and Yang's (2017) study indicated that co-author credits amounted to one at nine universities, less than one at three universities, and greater than one at 14 universities. Universities with low sums for co-author credits tended to have drastically low preferences for collaborative research. Researchers around the world are also highly aware of the authorship quota. Chinese universities have received criticism for only assigning credit to a single or lead author in their promotion of science engineering professors (Xu et al., 2018). There is growing concern that if credit is only granted to the lead authorship of a joint research project that is spread across humanities and social studies fields, collaboration between advisors and students will diminish. Despite these criticisms and concerns, Chinese universities are tightening their authorship requirements.

The authorship quota is also important in research evaluations; it is even more important when choosing where to publish. Lee and Yang (2017) analyzed research outputs evaluation criteria in the LIS field at 27 universities in South Korea. All South Korean universities implement evaluation methods that award credit points differently according to the journal. This has been going on for several decades. On average, 27 universities gave credit scores for the Korea Citation Index (KCI, https://www.kci.go.kr) 100, SSCI 249, and Scopus 142. Here, KCI is the citation index for South Korean journals produced by the NRF. According to them, SSCI credit is more than twice as high as for domestic journals, while the gap between points is larger among prestigious private universities. These universities are also implementing policies to differentiate credit according to the impact factor quartile of SSCI journals.

Many universities across the world also tend to allow high credit to highly-cited international journal publications. Nicholas et al. (2017) interviewed 116 ECR scientists, engineers, and social scientists from the United States, the United Kingdom, France, Spain, Poland, Malaysia, and China; their SSCI/SCI or Scopus journal-publishing preferences were clear. Furthermore, ECRs were confident that publishing in top international journals would enhance their careers.

Chinese ECRs are also a high priority in the publications of SSCI or SCI journals (Xu et al., 2018). Publishing in a world-renowned journal can increase the international influence of the research. Publication in such journals is also much more favourable for quantitative metrics research achievement evaluation and is very positive for individual recruitment, contract renewal, promotion, and tenure. These researchers are also making great efforts to publish in high-impact SSCI or SCI journals. The United States and China are the world leaders for these types of publications. Publishing in third-quartile journals or above is the next step for Chinese researchers who have already published many SSCI or SCI papers. This may also be the same context in which Chinese universities do not assign credit to Scopus journals.

Conversely, South Korean LIS authors do not seriously consider where to publish, especially about national journals. Some academic institutes are exceptions, but they allow the same credit when publishing in five leading domestic journals. The fact that there is little difference between the readers of domestic journals alleviates this consideration for South Korean authors. Choi and Yang (2018) confirmed that there were only 205 LIS professors in South Korea at the time of their study. This relatively small number made it unnecessary to worry about selecting South Korean journals. However, the competition among universities to publish in SSCI journals is becoming increasingly intense. There is also growing pressure for South Korean LIS authors to produce SSCI publications; this especially affects ECRs.

The choice of journal in which to publish is crucial for scholars across the globe. Researchers typically consider many factors aside from whether their university or research institution accepts these journals. Xu et al. (2018) conducted a survey on 11 factors considered in the journal selections of 14 Chinese ECRs. They responded that SSCI/SCI status, general prestige, and whether the journal was approved by their respective universities were highly important. In addition to other items, they did not value paper charges or innovative journal features. ECRs tend to target several journals before submitting to the most relevant. There are also cases in which a target scientific journal is established at the beginning of the project. Based on both academic publishing trends and previous research reviews, this study set out the following research questions:

1) What are the trends associated with academic journal publications and authorship for South Korean ECRs in LIS?
2) What difficulties are involved in journal publishing and authorship for research conducted by South Korean ECRs in LIS, and what are the proper solutions?

3. METHODS

This study targeted the recent journal publishing practices of active ECRs conducting research as South Korean LIS scholars. Thus, ECR journal publications in the recent five years from 2014 to 2018 were analyzed among 23 participants of an NRF-sponsored LIS project titled "The Young Researcher Program."

Despite intense competition, three of these scholars received more than one project, while only 19 ECRs were targeted. Specific ECRs were chosen because they presented more research output than other new researchers, thus appearing to be more interested in publishing and authorship and more likely to provide constructive feedback through interviews. This research also referred to journal publications from ECRs who contracted projects as long as a few decades ago. The Korea Researcher Information (KRI, https://www.kri.go.kr/kri2) site was also searched to determine whether LIS-related tasks were selected for the "Program for Emerging Research Fellows" project (now called the young researcher program) before and after 2000. Two projects were found, one from 1998 and the other from 2002. Projects older than these could no longer be found.

The project details, research achievements, and personal profiles of the ECRs were easily obtained from the NRF's KRI website. Table 1 shows the characteristics of the 19 ECRs targeted in this study. Except for one, all were working at colleges or universities. Each ECR had earned a PhD. Six held PhDs from foreign universities (31.58%). Seventeen (89.48%) were in the LIS field, while one was in computer science and the other was in communication studies.

However, the KRI site only showed the number of authors for collaborative studies. Thus, this study pooled the roles of lead, corresponding, and non-lead authorship using two sites (i.e., DBpia at https://www.dbpia.co.kr/ and Google Scholar at https://scholar.google.co.kr). Excluding conference proceedings and books, individual journal publications were collected up to five years prior to the year the project was granted. Rather than finding the journal impact factor, this study also collected impact factor quartile rankings provided by Journal Citation Reports (JCR). This is because journal impact factor is a proxy for journal quality level rather than a precise measure of quality (i.e., it is more appropriate to use the impact factor quartile ranking of journals within the LIS field).

The NRF examines domestic journals in South Korea for evaluation criteria (e.g., regular publishing, strict peer review, and research ethics compliance). Currently, 2,024 journals are listed by KCI; South Korean academic institutes assign credit to authors who publish in this journal. KCI also provides citation counts. However, the collection period was too short for use in this study. For non-SSCI/KCI journals, this study obtained necessary information from individual journal websites.

The trend in ECR journal publishing and authorship can be determined by analyzing metric data (e.g., journal publication productivity and authorship). However, it is difficult to identify the background to and cause of this trend. The publication trend is the cumulative result of the selection and decision processes of ECRs according to each publishing journal. It is, therefore, possible to identify the causes of any related publishing trends by gathering and analyzing the selection and judgment criteria of publishing journals selected by ECRs. These interviews were also useful in determining any related authorship practices or difficulties. In summary, this study conducted a quantitative metric analysis as its primary research method, while a qualitative data analysis was used for support. The specific interview questions were as follows:

Table 1. Details of selected early career researchers

Category	Subcategory	Details	Persons (%)	Subtotal
Affiliation	College/university	Department of library & information science	16 (84.22)	18 (94.74)
		Department of culture, tourism, & contents	1 (5.26)	
		College of general education	1 (5.26)	
	Research institute		1 (5.26)	1 (5.26)
PhD	Domestic		13 (68.42)	13 (68.42)
	Foreign		6 (31.58)	6 (31.58)
Subject field	Library & information science	Information service	4 (21.05)	17 (89.48)
		Library management	4 (21.05)	
		Cataloging & classification	3 (15.80)	
		Bibliography	2 (10.53)	
		Records management & archives	2 (10.53)	
		Digital library	1 (5.26)	
		Information science	1 (5.26)	
	Computer science	Machine learning	1 (5.26)	1 (5.26)
	Communication studies	Organizational communication	1 (5.26)	1 (5.26)
Total			19 (100.00)	

#1) What influence did you have on the choice of journal in which your research was published?
#2) Was there pressure on you to publish in top-ranked journals?
#3a) Does your department, university, or funder have a set of formal authorship guidelines for assigning author roles? If not, how do you decide?
#3b) Have you ever experienced an inappropriate authorship assignment or listed a non-contributing advisor/senior as an author?
#4) What difficulties have you faced in journal publishing and authorship? What factors are you dissatisfied with within these areas?

To identify #1, this study derived 10 key factors that could influence journal choice and created ten questions based on these. Nine of the 11 factors proposed by Xu et al. (2018) were used as they were, and one was newly created. Thus the newly created question of this study was "Did you choose the journal because it was an SSCI or KCI journal?" The two questions that Xu et al. (2018) presented but were not adopted in this study are as follows: Is the journal indexed in WoS? Is the journal approved by your university? Their two questions were not only partially overlapping, but also seemed difficult to answer quickly when they are easily understood by interviewees. For these finalized 10 questions, this study asked for answers with a 5-point Likert scale. The findings of #1 are described in Section 4.2.1. Unlike #1, questions from #2 to #4 were open-ended questions that respondents could freely answer, so a transcript of the interview was taken down in note form. Notable contents from the #2, #3a, #3b, and #4 interview responses were taken and summarized by the researcher.

Journal publication productivity and authorship was conducted using articles published by 19 ECRs. On the other hand, the interviews for #1 to #4 were performed only on six ECRs who had published in both domestic and SSCI journals. Interviews with these six ECRs provided a clearer picture of the practices and difficulties of SSCI journal publishing as well as domestic journals.

A telephone interview was conducted by the researcher in August 2018.These interviews were set up in advance and conducted according to appointments. According to the interviewees, the length of interview time was somewhat different, but the average time per interview was about 30 minutes. These ECRs comprehensively responded to the interview questions. They were aware that their identities would not be revealed and thus did not hesitate in revealing difficulties, complaints, and honest opinions. The interview results for #1 to #4 were given in Section 4.2.

4. RESULTS

4.1. Current Journal Publishing and Authorship Practices

Over the five-year period examined in this study, four outstanding ECRs published 17 or 18 papers, while another four published five or fewer. This study also searched the personal websites of the ECRs (which did not contain papers) and discovered additional research achievements (e.g., conference proceedings and books).

Table 2 shows the status of the articles published. The 19 studied ECRs published between 0 and 18 articles over the five-year period prior to the NRF project. They published a total of 194 articles (an average of 10.21), or two articles per person annually. These figures are higher than the average annual article counts for ECRs who earned NRF projects in 1998 and 2000.

Table 3 shows publications according to indexed database. The ECRs produced 166 domestic papers, accounting for 85.57% of the total. In addition, 28 papers were published in international journals (14.43%); of these, 25 were SSCI papers (12.89% of the overall total). When converted to an annual per-person statistic, SSCI papers were published at a rate of 0.26. In fact, six authors published between one and five SSCI papers. An analysis of each article revealed that 143 KCI papers were only published in some specific journals rather than being evenly distributed across all journals. A total of 116 papers were published in the top five KCI journals, accounting for 81.12% of all KCI papers. On the other hand, the 25 SSCI papers were evenly distributed across many journals. Of these, 12 were published in different journals.

Table 2. Summary of published articles

2014-2018 project						1998-2002 project			
Persons	Min	Max	Total	Average	Average per year	Persons	Total	Average	Average per year
19	0	18	194	10.21	2.04	2	6	3	0.6

Table 3. Article publishing status as classified by indexed database

Category	Details	Quantity	%	Subtotal	
				Publication count	%
Domestic	KCI	143	73.71	166	85.57
	Non-KCI	18	9.28		
	Others	5	2.58		
International	SSCI	25[a)]	12.89	28	14.43
	Scopus	1	0.52		
	Others	2	1.03		
Total		194	100	194	100

KCI, Korea Citation Index; SSCI, Social Science Citation Index.
[a)] SSCI articles produced by six authors who had written between 1 and 5.

Table 4. Distribution of author numbers

No. of authors	Publication count	%	Subtotal	% of subtotal
1	115	59.28	115	59.28
2	46	23.71	79	40.72
3	27	13.92		
4	4	2.06		
5	1	0.52		
≥6	1	0.52		
Total	194	100	194	100

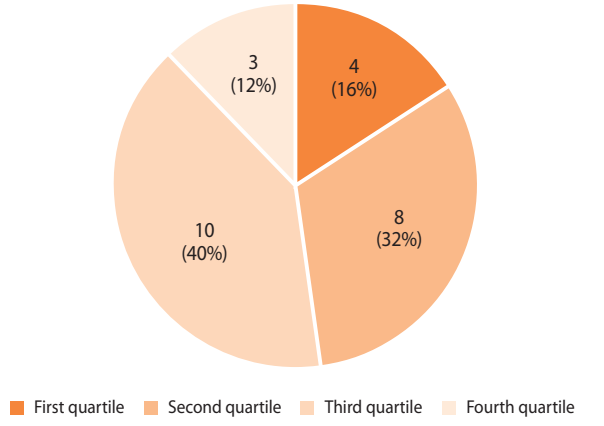


Fig. 1. Quartile ranking distribution of Social Science Citation Index-indexed papers.

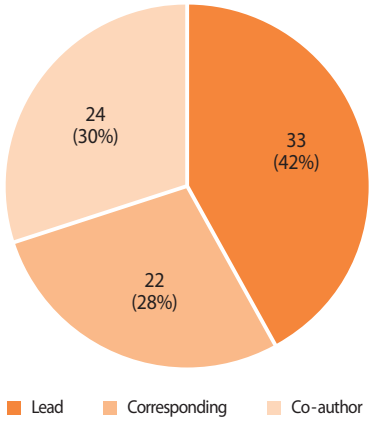


Fig. 2. Published papers classified by authorship type.

Fig. 1 shows the impact factor quartile ranking distribution of the SSCI journals in which the ECRs published. The number of titles of each quartile was 4, 8, 10, and 3, respectively; the quartile with the highest frequency was Q3. Q3 and Q2 combined for a total of 18, accounting for 72% of all SSCI publications. This means that the majority of SSCI papers written by South Korean ECRs in LIS are of an intermediate grade. Nevertheless, the overall quartile grade of the papers was not low (almost half of the 25 SSCI papers were Q1 and Q2).

Table 4 shows the number of authors for all 194 publications. Of this total, 115 (59.28%) were single studies, while 79 were joint studies. For joint studies, 46 papers were written by two authors, while only two had more than five authors. Fig. 2 shows the categorization of 79 collaborative studies conducted by ECRs according to author roles. As a result, there were 33 lead authorship papers, 22 corresponding authorship papers, and 24 non-lead authorship papers. A total of 70% of all collaborative studies were from main authors (e.g., lead or corresponding authorship).

4.2. Journal Choice Practices and Difficulties Identified Through Interviews

4.2.1. Journal Choice

With the journal titles published alone, it is difficult to know exactly what criteria the ECRs used when choosing journals. This study identified these criteria by conducting interviews with the six studied ECRs who published in both domestic and SSCI journals. Here, almost the same selection criteria from Xu et al. (2018) were used. As shown in Table 5, both results revealed an analogous journal selection policy that first involves publishing in SSCI journals and those approved by universities. Page charges and innovative features had little impact on journal selection. South Korean ECRs appear to be more concerned with high impact factors, topic relevance, and specialist audiences than those researched by Xu et al. (2018).

4.2.2. Pressure to Publish in Top-ranked Journals

Random letters were assigned also to the six interviewees.

During the interview, all six ECRs indicated that they felt pressure from their respective departments/universities to publish in SSCI journals (even the top-ranked).

Author A said he was working on publishing in a Q1 SSCI journal and that his university was planning research achievement standards that would only allocate credit for Q1 journals. Despite not being Q1 journals, he said that some specific journals were highly reputable to specialist readers and that he was sorry he was unable to publish in them.

Author B had similar opinions, although she was highly interested in school libraries and had published in an authoritative school library journal with a small readership (this was not an SSCI journal). She earned very little credit for her paper (even less than that given to KCI journals). She admitted that she was very disappointed. She has since decided to publish the paper in a university-approved SSCI journal instead of the specialized journal in which she first wanted to publish.

Of the difficulties related to journal-publishing choice, author C said that there was too much room to choose, but there was a great deal of pressure to publish in top journals. At her department/university, only two LIS journals listed by KCI were approved as national journals. She complained about her department/university, which did not accept all eight LIS journals listed by KCI for credit. In this situation, she explained that researchers will inevitably wait on SSCI/SCI/A&HCI journal publishing instead of submitting to the few LIS journals listed by KCI that were also approved by her department/university. She also had suspicions that her department/university deliberately insisted on this policy to prevent authors from frequently submitting to KCI journals.

4.2.3. Authorship Guidelines and Inappropriate Authorship Practices

All interview participants said they had no official principles or guidelines for authoring (e.g., those stipulating lead, corresponding, or non-lead authorship). Rather, the lead or corresponding author was generally chosen among all collaborators. However, more specific criteria have now been set; it would help if a documented research management guide was published. For example, the person first proposing the research idea, leading the research initiative, and contributing more than half of the manuscript should be credited with lead authorship.

Except for one, the interviewed ECRs were satisfied with role assignments for multi-authored works. Author C actively participated in publishing research but was not listed as a lead or corresponding author. She complained that she had listed a senior professor who did not participate in writing the paper (a so-called ghost author) as a lead author. The five other ECRs said they had never been pressured to list a senior/advisor who did not contribute to the paper as a co-author.

Table 5. Researchers' responses to factors considered in deciding where to publish

Factor	Importance[a]	
	This study (Korean 6, 2018)	Xu et al. (Chinese 14, 2018)
Indexed in WoS (SCI, SSCI, A&HCI)	5[b]	5
Journal approved by university	5[b]	5
High impact factor	4	2
Most relevant to the field	4	2
General prestige	4	5
Fast manuscript processing	4	3
Covers specialist audience	3	1
Open access	2	1
High level of peer review	2	3
No page charges	1	0
Innovative features	1	0

WoS, Web of Science; SCI, Science Citation Index; SSCI, Social Science Citation Index; A&HCI, Arts & Humanities Citation Index.
[a]The average scores provided by interviewees who rated criteria on a graduated scale from 1=not important to 5=extremely important.
[b]The question was changed to 'SSCI or Korea Citation Index?' (Refer 3. METHODS).

4.2.4. Additional Comments on Publishing or Authorship

All six ECRs experienced stress during the journal publishing process. Author C said competition between universities raised the research achievement standards for each department/university every few years. Even when a PhD was employed in the same department, the volume and quality standards of research output required for different tenure promotions depended on the promotion policy regarding the employed year of the employee. ECRs are required to publish in journals that meet their own standards. Rather than pushing journal publication on a tenured professor, it is often necessary to boost the quality and quantity of journal publishing required for the promotion of assistant or associate professors. Due to this academic environment, she stated that she always considered a journal that was likely to be accepted among the university-approved KCI or SSCI journals rather than one simply regarded as suitable for the manuscript.

On the other hand, five ECRs excepting author F pointed out the limitations of peer review as a difficulty in journal publishing. These ECRs were confused because referees demanded that the contents or methods listed in the manuscript be considerably revised without full understanding. This was because there were not enough referees in the sub-discipline; such reviews were the results of adjudicator screening. The interviewees also asserted that specialists who were familiar with their sub-disciplines should be supplemented to achieve fair and constructive review feedback.

5. DISCUSSION AND CONCLUSIONS

This study gathered journal publishing data from 19 South Korean ECRs in LIS. Subjects were chosen to participate in the NRF's New Researcher Program from 2014 to 2018. This study analyzed the productivity, journals, authorship, and pressures associated with their publishing activities. Six ECRs were also interviewed to ascertain their experiences and opinions related to journal publishing and authorship.

The status of journal publishing for ECRs was generally positive. The 19 studied ECRs published a total of 194 papers over the last five years. Their level of productivity was thus considered good. Domestic papers accounted for 166 (85.57%) of the total. This number was overwhelmingly higher than the number of international papers (28, 14.43%) (Table 3). South Korean researchers typically write papers in their native language and publish the results in a domestic journal. However, such studies are obviously limited in that they are only shared with readers proficient in Korean. Publishing in domestic journals only partly accomplishes the original purpose of publishing in journals that contribute to the development of science by widely spreading research. In this context, it is understandable that South Korean departments/universities do not award high credit to domestic journal publications. Meanwhile, international journal publications (especially SSCI journals) receive high levels of credit and are encouraged (average credits among 27 universities: KCI, 100; SSCI, 249; Scopus, 142) (Lee & Yang, 2017). SSCI journal papers thus receive approximately 2.5 times the credit of KCI publications. Even at this writer's university, SSCI journal publications are given three times the amount of credit as KCI papers (Sejong University, 2018).

The difficulties for ECRs in publishing in academic journals became more apparent through the interview process. Departments/universities seemed to pressure these ECRs to publish in Q1 SSCI journals. However, this study has already presented that among the 25 SSCI papers written by the examined ECRs, there were only four Q1 papers (16%) (Fig. 1). One ECR stated that his university was preparing a rigorous promotion screening standard that would only reflect Q1 journal publications and the remaining journals would not be recognized. He showed considerable shock and dissatisfaction with these changes. There are several reasons that his university insists on this policy. First, major South Korean universities are competing for positive reputations and rankings. It is possible that obtaining a university rating in the top 10% of papers published (Leiden Ranking, 2018) according to the Leiden ranking may provide considerable prestige. South Korean universities seem to have already found a way to add criteria that imitate the Leiden ranking when conducting faculty research evaluations. Several major universities (including mine) have subdivided the SSCI according to the impact factor quartiles and have begun to assign credits accordingly (Sejong University, 2018). In this way, universities seem to believe that differentiating credits for each quartile has effects that accompany qualitative paper evaluations. Since South Korean universities have used quantitative assessments such as SSCI or KCI paper counts to assess faculty research achievements for decades, it may now be easier for them to make these decisions using the quartiles directly. Thus, the active use of quartile ratings attracts attention because it allows university headquarters to take the initiative in evaluating the research achievements of all professors instead of the senior faculty members of each department only.

Rather than choosing a journal that fit their manuscript topic, most ECRs said they often published in journals with high evaluation credits in which they were likely to pass referee reviews. These were selected among the journals approved by

their departments/universities. Thus, they rarely published in journals with low or no credits. For example, if a journal is enthusiastically read, but only by a narrower readership in the sub-discipline, the impact factor/credit is often low. Some ECRs thus said that even if a journal was suitable for their manuscript, it was unwise to publish in it. Although all researchers (including ECRs) are decisive for immediate research evaluation, it has long been necessary to consider publishing in journals (e.g., sub-discipline journals) that are useful for career advancement (Nisonger & Davis, 2005). Referring to the above, the South Korean university headquarters directly involved in the journal publishing choices of their faculty and which have initiative in the evaluation of their research achievements may invite side effects that will distort the landscape of the academic publishing industry. Researchers, universities, and institutes seeking rankings all seem to require serious improvements in conducting their evaluations of journal articles for long-term academic advancement.

On the other hand, South Korean ECRs seem to have no major authorship problems. Of the total papers, the ratio of single to multi-authored papers was 6:4 (Fig. 2). As not many joint studies were conducted, little conflict was involved in authorship role allocation. In joint studies, for instance, 70% of Korean ECRs have lead or corresponding authorship, while only 30% had simple co-authorship. The interviews also revealed that authors were not disadvantaged by receiving a contributor's role while acting as junior researchers. Only one ECR identified abusive practices in which a listed author did not directly participate in the paper.

Collaborative research is gaining global popularity. It is thus notable that South Korean ECRs in LIS do not show related publishing patterns. Collaboration can encourage author productivity and enhance paper quality not only in science, technology, and medicine but also in social sciences (Bahr & Zemon, 2000; Lee & Bak, 2016). It is therefore worth noting that there is another reason that ECRs were not active in collaborative research. That is, single authorship is recognized as 100% of the total score, but there is concern that departments/universities may have low recognition rates for co-authors. Reports have indicated that three universities award co-author credits below 100% (Lee & Yang, 2017). A university with such a policy may operate with the suspicion that some authors are listed without contributing. However, modern practice involves various digital traces that remain after the collaborative research has concluded. This is because documents are shared by e-mail or through personal cloud storage (e.g., Dropbox) during the process. It is therefore unlikely that unethical social norms (e.g., free-riding or senior preference) will arise. It is thus important to point out that it is an antiquated practice to lower co-authorship credits.

Authorship guidelines seem highly necessary for dissolving the doubts of these universities while encouraging healthy collaboration. Such guidelines can also be referred to when assigning participatory roles; it has been confirmed that this needs to be specified in future projects. The International Committee of Medical Journal Editors and the Committee on Publications Ethics provide good examples of who should be listed as an author (Brand et al., 2015; International Committee of Medical Journal Editors, 2018). All six ECRs participating in this study's interviews were also very helpful in discussing authorship guidelines and helping to achieve collaboration. These ECRs held the common opinion that collaboration with overseas researchers should be encouraged to actively pursue SSCI publishing.

In the South Korea LIS field there are not enough research projects being carried out through joint research. The scientific community is not large enough to facilitate such collaboration. Less credit is thus given for these efforts. In the case of LIS, a joint study between researcher and librarian, not between researchers, would be particularly practical. In South Korea, however, little research has been done. Future studies will likely reveal the cause of this problem and appropriate countermeasures. More comprehensive and meaningful implications also can emerge if information on the journal publishing practices of senior researchers is also analyzed and combined with this study's results.

REFERENCES

Adkins, D., & Budd, J. (2006). Scholarly productivity of US LIS faculty. *Library & Information Science Research*, 28(3), 374-389.

Bahr, A. H., & Zemon, M. (2000). Collaborative authorship in the journal literature: Perspectives for academic librarians who wish to publish. *College & Research Libraries*, 61(5), 410-419.

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, 28(2), 151-155.

Choi, E. J., & Yang, K.-D (2018). A bibliometric analysis of library and information science research in Korea: 2001-2016. *Discourse and Policy in Social Science*, 11(1), 55-86.

Chung, J. Y., & Park, J. H. (2011). Analysis of the trends in the field studies of library and information science in Korea.

Journal of Korean Library and Information Science Society, 42(2), 171-191.

Davarpanah, M., & Aslekia, S. (2008). A scientometric analysis of international LIS journals: Productivity and characteristics. *Scientometrics*, 77(1), 21-39.

Frandsen, T. F., & Nicolaisen, J. (2010). What is in a name? Credit assignment practices in different disciplines. *Journal of Informetrics*, 4(4), 608-617.

International Committee of Medical Journal Editors (2018). *Defining the role of authors and contributors*. Retrieved September 1, 2018 from http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html.

Larivière, V., Sugimoto, C. R., & Cronin, B. (2012). A bibliometric chronicling of library and information science's first hundred years. *Journal of the American Society for Information Science and Technology*, 63(5), 997-1016.

Lee, H.-K., & Yang, K. (2017). Comparative analysis of Korean universities' journal publication research performance evaluation standards. *Journal of Korean Library and Information Science Society*, 48(2), 295-322.

Lee, J., & Yang, K. (2011a). A bibliometric study of library and information science research in Korea. *Journal of the Korean Society for Library and Information Science*, 45(4), 53-76.

Lee, J., & Yang, K. (2011b). A bibliometric analysis of faculty research performance assessment methods. *Journal of the Korean Society for Information Management*, 28(4), 119-140.

Lee, J. W., & Bak, H. R. (2016). Characteristics of Korean researchers through bibliometric analysis of papers published in international LIS journals. *Journal of Korean Library and Information*, 47(1), 217-242.

Lee, S.-H., & You, B.-J. (2014). A study on the research trends of library and information science in Korea using S&T authority data. *Journal of the Korean Society for Library and Information Science*, 48(4), 377-399.

Leiden Ranking (2018). *Selection of universities*. Retrieved September 1, 2018 from http://www.leidenranking.com/information/universities.

Maciejovsky, B., Budescu, D. V., & Ariely, D. (2009). The researcher as a consumer of scientific publications: How do name-ordering conventions affect inferences about contribution credits? *Marketing Science*, 28(3), 589-598.

Mukherjee, B. (2010). Assessing Asian scholarly research in library and information science: A quantitative view as reflected in Web of Knowledge. *Journal of Academic Librarianship*, 36(1), 90-101.

Nicholas, D., Rodríguez-Bravo, B., Watkinson, A., Boukacem-Zeghmouri, C., Herman, E., Xu, J., . . . Świgoń, M. (2017). Early career researchers and their publishing and authorship practices. *Learned Publishing*, 30(3), 205-217.

Nisonger, T. E., & Davis, C. H. (2005). The perception of library and information science journals by LIS education deans and ARL library directors: A replication of the Kohl-Davis study. *Colleges & Research Libraries*, 66(4), 341-377.

Sejong University (2018). *Guidebook of research works*. Retrieved September 1, 2018 from http://rnd.sejong.ac.kr/contents/rnd/cor/study.html.

Shaw, D., & Vaughan, L. (2008). Publication and citation patterns among LIS faculty: Profiling a "typical professor". *Library & Information Science Research*, 30(1), 47-55.

Xu, J., Nicholas, D., Zeng, Y., Su, J., & Watkinson, A. (2018). Chinese early-career researchers' scholarly communication attitudes and behaviours: Changes observed in year two of a longitudinal study. *Journal of Scholarly Publishing*, 49(3), 320-344.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 17-28, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.2

Dengue-related Information Needs and Seeking Behavior of the General Public in Singapore

Shaheen Majid*
Nanyang Technological University, Singapore
E-mail: asmajid@ntu.edu.sg

Hu Ye
Nanyang Technological University, Singapore
E-mail: huye0001@e.ntu.edu.sg

Hui Yik Tan
Nanyang Technological University, Singapore
E-mail: huiy0005@e.ntu.edu.sg

Lin Xinying
Nanyang Technological University, Singapore
E-mail: xlin013@e.ntu.edu.sg

ABSTRACT
Dengue infection is becoming a serious global health threat. Public awareness is a pre-requisite for the successful implementation of dengue prevention programs. The main purpose of this study was to investigate dengue-related information needs and seeking behavior of the general public in Singapore. Some areas covered by this study were: importance of dengue-related information needs, preferred channels for seeking information, and respondents' perceptions of using dengue-related information. A questionnaire was used for data collection and 152 individuals participated in this study. Data analysis showed that the most sought after information concerned: dengue-related medicines, primary symptoms of dengue infection, and different possible treatments. The popular channels for seeking information were: websites of hospitals and other health agencies, the social media, television, and newspapers. Medical staff, such as doctors and nurses, were trusted for providing accurate information. Although credibility of social media was considered low, respondents were using it due to its easy accessibility. The findings of this study will be useful to government health departments in Singapore as well as in other countries suffering from dengue, hospitals, and public welfare agencies involved in public health awareness campaigns.

Keywords: dengue fever, information needs, information seeking, information quality, communication channels, Singapore

Open Access

Accepted date: March 08, 2019
Received date: January 27, 2019

***Corresponding Author:** Shaheen Majid
Associate Professor
Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, Singapore 637718
E-mail: asmajid@ntu.edu.sg

1. INTRODUCTION

In recent years, dengue has become a widespread worldwide health threat, particularly in Southeast Asia and Latin American countries. The estimated numbers of global annual dengue infections have reached 390 million. According to the World Health Organization, dengue is a mosquito-borne viral infection causing a severe flu-like illness, sometimes leading to potentially lethal complications (World Health Organization, 2019a). The incidence of dengue has increased 30-fold over the last 50 years. Up to 50 to 100 million infections are now estimated to occur annually in over 100 countries, putting almost one-half of the world's population at risk (World Health Organization, 2019b). However, it mostly affects Asian and Latin American countries and has become a major cause of hospitalization and deaths among children and adults. The dengue virus is transmitted through the female *Aedes aegypti* mosquito which acts as a vector (or transmitter). The dengue virus is passed on to humans through the bites of infective female *Aedes* mosquitos, which mainly acquire this virus from the blood of infected persons.

In spite of concerted efforts by multiple agencies, Singapore is still unable to completely wipe out dengue infections. It could partially be due to the reason that Singapore is a popular business and tourism destination and millions of travelers visit Singapore every year, some of whom may be carrying the dengue virus. In addition, hot and humid weather provides good breeding grounds for mosquitos. According to the National Environment Agency, 1,183 dengue infection cases were recorded in Singapore during the last three months of 2018, an increase of 63.6% as compared to the previous quarter (July to September 2018) (National Environment Agency, 2018).

Different government agencies in Singapore use a variety of methods to create awareness about dengue infection through posters, handouts, newspaper articles and advertisements, radio and television interviews, talks and documentaries, public seminars, and various other community awareness programs. However, these efforts could be more fruitful if information dissemination to members of the general public were based on their information needs and information seeking behavior. The main objective of this study was to investigate dengue-related information needs of the general public, their preferred information seeking channels, and their perceptions of the information quality disseminated through different communication channels.

2. LITERATURE REVIEW

Dengue is one of the most rapidly spreading mosquito-borne diseases in the world. It is transmitted by the *Aedes* mosquitoes and mainly found in tropical and semi-tropical urban areas of the world (Boonchutima, Kachentawa, Limpavithayakul, & Prachansri, 2017; Elsinga et al., 2018). The major reasons for its speedy spread are dense human populations, frequent international travelers, and poor vector control (Bhatt et al., 2013). The dengue infection may result in a series of clinical symptoms usually appearing 4 to 7 days after the mosquito bite, which include fever, skin rashes, intense headache, vomiting, and bleeding from nose or gums (Singapore Ministry of Health, 2018). In severe cases, dengue can potentially trigger lethal complications such as plasma leakage, hemorrhagic manifestations, organ impairment and shock, and even death (Singapore Ministry of Health, 2018; World Health Organization, 2019a).

Singapore, with a tropical rainforest forest climate, provides a conducive environment for the spread of all four serotypes of dengue virus (i.e., DEN-1, DEN-2, DEN-3, and DEN-4) and out of these DEN-2 and DEN-3 are considered more dangerous as they can cause secondary dengue infections (Singapore Ministry of Health, 2018). The lack of distinctive seasons, heavy and frequent rainfall, and a highly urbanized environment provide ideal conditions for *Aedes* mosquito habitation in Singapore (Lee et al., 2012; Seltenrich, 2016). Every year, the highest number of dengue cases in Singapore are recorded during July and August months due to high mean temperature (Viennet, Ritchie, Williams, Faddy, & Harley, 2016).

In 2013, there was a record 22,248 reported dengue cases in Singapore, with 842 cases in a single week. In 2016, a total of 13,115 dengue cases was reported in Singapore and 9 of these resulted in deaths. Comparatively, 2018 was a better year as there were a total of 1,282 dengue infection cases with only 5 deaths (Singapore Ministry of Health, 2018). In addition to other factors, lack of adequate dengue prevention awareness and insufficient public engagement were the reasons for the resurgence of dengue infections in Singapore (Viennet et al., 2016). It is in spite of the fact that several government agencies actively participate in the dengue prevention awareness campaigns and use a variety of communication channels to spread the message to different segments of society (Rajarethinam et al., 2018).

For implementing an effective disease awareness program, it is desirable to first adequately understand the knowledge level, information needs, and information seeking behavior of the general public. Firdous et al. (2017) stressed that more

practical health education programs were needed to create awareness and a positive attitude among the community members for effectively preventing and controlling dengue infections. A study by Elsinga et al. (2018) in Venezuela found that preventive practices were associated with better knowledge of dengue symptoms and transmission routes. It was also revealed that dengue-related knowledge was associated with exposure to various information sources. A study in Lima, Peru revealed that low knowledge among the respondents was mainly due to lack of access to appropriate information about dengue virus transmission modes, and the desired measures for its control (Cabrera et al., 2016).

Children are usually more vulnerable to dengue infection, and therefore they need to be knowledgeable about the symptoms and preventative measures. A study involving 601 school students in Islamabad, Pakistan investigated dengue-related knowledge of students exposed to different dengue awareness campaigns. It was found that two-thirds of the students had poor knowledge about dengue, particularly about the spread of its virus (Javed, Ghazanfar, & Naseem, 2018). A majority of the students (72.9%) reported acquiring dengue-related information from television and radio. In addition, 44.6% of the students also acquired this knowledge through the dengue awareness campaigns run by their respective schools. A study of 362 students in Makkah, Saudi Arabia concluded that school-based awareness campaigns and social mobilization can help raise awareness about dengue infection and this knowledge can be translated into sound preventative practices (Alhazmi et al., 2016). Siriwardana and Samarasinghe (2018) suggest that school teachers in Sri Lanka can play a vital role in transferring dengue-related knowledge to students in particular and the community in general. Their findings, based on focus groups and semi-structured interviews, revealed that school teachers possess good knowledge about the spread and symptoms of dengue infection. It was concluded that school teachers, due to their positive attitude, knowledge of dengue infection, and social status in the society, can be used for preventing the spread of this disease.

Basically the spread of dengue virus can be prevented through controlling the population of *Aedes* mosquitoes. Community education, awareness, and cooperation can help reduce the spread of this disease (Harish et al., 2018). In order to successfully reduce dengue infection cases, a comprehensive citizen education and awareness program could be useful (Majid & Rahmat, 2013). Such awareness programs and campaigns can help the general public to remain vigilant and take the necessary precautions to protect themselves from dengue infections. Ooi, Goh, and Gubler (2006) reported that the perceived severity and community awareness of dengue infection can influence the effectiveness of disease prevention and control measures. A study investigating the impact of information, education, and communication on dengue awareness in India showed a significant improvement in dengue-related knowledge of urban households (Nivedita, 2016). Similarly, in Thailand it was found that although information from awareness campaigns regarding malaria and dengue were reaching the target population, certain specific knowledge gaps still exist (Brusich et al., 2015).

A study in Bangladesh by Chowdhury, Haque, and Meyur (2013) found that a close coordination between dengue patients and primary healthcare providers was useful in successfully implementing dengue control measures. A telephone survey of 1,050 individuals in Malaysia showed that, during the H1N1 outbreak, a majority of the people wanted to receive information about virus prevention and treatment through health authorities and mass media (Wong & Sam, 2010). A study by Majid and Rahmat (2013) during the H1N1 outbreak in Singapore found that the most sought after information concerned the disease symptoms, causes of the infection, preventative measures, and possible treatments. During the outbreak of H7N9 virus in Guangzhou, China people were generally more interested in receiving information about disease prevention (75.31%), the current epidemic situation (71.86%), and vaccination availability (63.10%) (Li et al., 2014). On the whole, it appeared that during epidemics a considerable proportion of the general public usually prefer receiving information about the disease symptoms, preventive measures, current rate of infections, and possible treatments.

In addition to understanding the information needs of the general public for epidemic diseases, it is equally important to know about their preferences for different information sources as well as their information seeking behavior. Voeten et al. (2009) reported that disease-related information seeking behavior may determine people's knowledge of diseases, which in turn might influence their health beliefs and precautionary actions. Some other studies also suggest that identification of the information seeking behavior of the general public towards certain diseases is essential for developing customized information distribution strategies to better cater to their information needs (Li et al., 2014; Wong & Sam, 2010). Harish et al. (2018) interviewed 195 parents in Bangalore, India whose children were hospitalized for dengue infection, to assess their knowledge regarding dengue fever, its transmission, possible infection complications, and the desired preventive measures. It was found that television, followed by radio and newspapers, were the popular channels for getting dengue-related information. They concluded that improved knowledge about dengue preventive measures can

ultimately help reduce the transmission of dengue virus in a community.

A study by Wong and Sam (2010) found that visual media, such as television, are more powerful in promoting disease awareness than written media because visual media are more understandable to many people. In Pakistan, Siddiqui, Ghazal, Bibi, Ahmed, and Sajjad (2016) also found that television served as a better communication channel in disseminating dengue-related information than radio and newspapers due to its popularity across different socioeconomic groups in the country. Another study in Pakistan by Hassan, Khail, Waris, Alam, and Marwat (2017) revealed that television and radio were the most preferred sources for seeking dengue-related information. Similar conclusions were also drawn by Yboa and Labrague (2013), who investigated the preferred resources for dengue information among rural population in the Philippines. A study involving 7,772 individuals from 25 provinces of Thailand stressed the role of media in educating and reminding Thai people about the prevention and control of dengue infection (Boonchutima et al., 2017). Studies in certain other countries, such as Malaysia (Hanim et al., 2017), Singapore (Majid & Rahmat, 2013), China (Li et al., 2014), and Brazil (Alves et al., 2016) have also reported the vital role played by mass media and online sources in creating awareness about dengue and other mosquito borne diseases. On the whole, it appeared that for dengue-related information, a majority of individuals preferred receiving information through mass media channels.

The literature review revealed that a majority of the previous studies have mainly focused on the knowledge, attitudes, and practices aspect of dengue infections. Only a limited number of studies are available on dengue-related information needs and information seeking behavior. The purpose of this study was to bridge this knowledge gap and investigate the dengue-related information needs of the general public in Singapore and their preference for different information sources and communication channels. Some areas covered by this study were: frequently sought after dengue-related information, preferred information sources, use of different communication channels, quality of information accessible through different channels, and respondents' perceptions of utilizing dengue-related information. The findings of this study will be useful to health information communicators, hospitals, government health departments, social welfare departments, and other agencies involved in public health and safety in Singapore as well as in other countries. They can also use these findings to assess the effectiveness of their existing dengue awareness campaigns as well as use this knowledge for developing future health awareness strategies.

3. METHOD

This study used a quantitative approach of a questionnaire survey for data collection. Initially content analysis of dengue-related brochures, posters, and other sources was conducted to understand the type of information usually disseminated to the general public. This information was used for designing the survey instrument. The questionnaires of some previous studies on similar topics were also consulted to learn about the type of questions asked and the measurements used by them (Majid & Rahmat, 2013; Li et al., 2014; Wong & Sam, 2010).

The survey questionnaire consisted of four sections and the first section asked respondents about the importance of different dengue-related information needs. A 5-point Likert scale was used to record their responses, where 1 represented the "least important" and 5 the "most important." The next question in this section asked the respondents about the information and communication channels used by them for meeting these information needs. The purpose was to map the information needs of the respondents with their preferred information sources and channels. The information sources/channels included in the questionnaire were: mass media channels (i.e., television, radio, newspapers), Internet sources (i.e., websites of hospitals and other health agencies, social media, general websites including Wikipedia), and other sources which can provide dengue-related information (i.e., healthcare providers such as doctors and nurses; family members, friends, and co-workers; printed posters/brochures). The second section of the questionnaire asked the respondents about their perceptions of the quality attributes of information acquired through these sources or channels. The attributes listed in the questionnaire were: accuracy, timeliness, accessibility, understandability, and information adequacy. The third section of the questionnaire asked about respondents' perceptions of using dengue-related information. A 5-point agreement scale was used, where 1 was "strongly disagree" while 5 was "strongly agree." The last section of the questionnaire sought demographic information on the respondents such as their gender, age group, education, and if they or any of their family members have ever suffered from dengue infection. The study and its questionnaire were approved by the institutional review board of Nanyang Technological University, Singapore (C1201617S2-005).

The questionnaire was pre-tested on six individuals to determine if the language of the questions and their measurements were appropriate and easy to understand. Based on their feedback, some of the jargon used in the questionnaire such as virus types (e.g., DEN-1, DEN-2) and the technical name of dengue mosquitos (i.e., *Aedes aegypti*) were removed.

The pre-tested questionnaire was used for data collection from different geographical zones of Singapore. The questionnaire was distributed at places where people were likely to gather and have enough time to complete the questionnaire such as libraries, student activity centers, and food courts. A total of 152 individuals participated in this study.

4. RESULTS

4.1. Demographic Profile of the Respondents

Out of the 152 participants, 55.9% were male and 44.1% female (Table 1). The age groups of the respondents were not quite well-distributed, partly due to the convenience sampling method. Sixty-four (42.1%) of the respondents were from the age group of 21 to 30 years old, and 83 (54.6%) belonged to the age group of 31 to 40 years, while only 5 (3.3%) were older than 40 years. However, the ethnicity representation was close to the national demographic distribution. There were 114 (75%) Chinese participants, followed by Malays (13% or 8.6%), and Indians (7 or 4.6%). The remaining 18 (11.8%) respondents belonged to other nationalities and ethnic groups.

The majority (58.6%) of the respondents possessed a bachelor's degree while another 28.9% were holding a master's degree. Once again this skewed distribution was due to convenience sampling where data was also collected from different institutions of higher education.

Table 1. Demographic attributes of the respondents (n=152)

Demographic data		Frequency	Percent
Sex	Male	85	55.9
	Female	67	44.1
Age (yr)	21–30	64	42.1
	31–40	83	54.6
	41–50	3	2.0
	More than 50	2	1.3
Race	Chinese	114	75.0
	Malay	13	8.6
	Indian	7	4.6
	Other races	18	11.8
Education	Post-secondary or less	19	12.5
	Bachelor's degree	89	58.6
	Master's degree	44	28.9

4.2. Dengue-related Information Needs

The respondents were asked to indicate the importance of various dengue-related information needs, using a 5-point Likert scale, where 1 represented the "least important" and 5 the "most important" (Table 2). The top three most important dengue-related information needs were: availability of dengue-related medicines and vaccines in Singapore (mean score 4.24); primary symptoms of dengue infection (mean score 4.18); and different treatment options available for dengue infection (mean score 4.14). Some previous studies, although either on H1N1 or H7H9 epidemics, also showed that these were the most important information needs (Li et al., 2014; Majid & Rahmat, 2013; Wong & Sam, 2010). It appeared that for mosquito-borne diseases, the general public usually wish to receive information about the disease symptoms, different available treatments and vaccines, preventative measures, and the current epidemic situation.

On the contrary, the three least important information needs were: the origin of dengue virus (mean score 3.07),

Table 2. Importance of dengue-related information needs (n=152)

Ranking	Information needs	Importance level	
		Mean score (1–5)	Standard deviation
1	Availability of dengue-related medicines and vaccines in Singapore	4.24	0.94
2	Primary symptoms of dengue infection	4.18	0.85
3	Different treatment options for dengue	4.14	0.90
4	Dengue cluster areas in Singapore	4.01	0.94
5	Current dengue status (e.g., no. of infected cases) in my neighborhood	4.01	1.02
6	Potential breeding sites in homes or public areas	3.92	0.99
7	Vulnerable groups and their risk levels	3.81	0.93
8	Availability of protection products (e.g., insect repellent)	3.74	1.06
9	Death rate due to dengue	3.72	0.97
10	Peak biting time in a day of dengue mosquitoes	3.69	1.08
11	Incubation period of dengue virus	3.39	1.04
12	Origin of dengue virus	3.07	1.21

incubation period of dengue virus (mean score 3.39), and the peak biting time in a day of dengue mosquitoes (mean score 3.69). However, it was worth noting that the mean scores for all the listed information needs were more than 3, indicating that the respondents were interested in receiving information on all possible aspects of dengue infection.

4.3. Preferred Communication Channels

The respondents were asked about the information sources and communication channels used by them for seeking dengue-related information. The purpose was to map respondents' information needs with their preferred information sources and channels for seeking the needed information. Based on the top three most preferred channels for acquiring the needed information, it was found that "websites of hospitals and other health agencies" and "social media" were used to satisfy all the listed information needs (Table 3). General websites, including Wikipedia, were used by the respondents to get information about "primary symptoms of dengue infection," "incubation period of dengue virus," "the origin of dengue virus," and "peak biting time in a day of dengue mosquitoes."

Television was among the top three preferred channels for seeking information about "vulnerable groups and their risk levels," and "potential breeding sites in homes or public places." Newspapers were among the top three preferred channels for getting information about "dengue cluster areas in Singapore," and "current dengue status in the neighborhood." Finally, medical staff, such as doctors and nurses, were preferred for seeking information about "different treatment options available for dengue treatment," and "availability of dengue-related medicines and vaccines in Singapore."

Table 4 shows a ranking of the top three most preferred communication channels for satisfying each dengue-related information need. It was worth noting that out of 12 listed information needs, different Internet-based platforms were the first choice for 10 information needs. Television and newspapers were the first choice for only one information need each. It appeared that the participants were heavily dependent on Internet sources for seeking dengue-related information, followed by mass media (TV and newspapers). These findings are in line with some previous studies which showed that online information sources (Li et al., 2014) and mass media (Boonchutima et al., 2017; Harish et al., 2018; Javed et al., 2018) were the most heavily used sources for seeking information about dengue infection.

A summary of the top three preferred sources for 12 dengue-related information needs is given in Table 5. Websites of hospitals and other health agencies as well as social media were considered suitable for meeting 100% of dengue-related information needs. General websites were expected to meet one-third of dengue-related information needs. However, it was worth noting that medical staff (doctors, nurses, technicians,

Table 3. Preferred communication channels for seeking dengue-related information (multiple response)

Information needs	Mass media			Internet			Human sources		Print
	TV	Radio	News-papers	Websites of hospitals and others	Social media	General websites	Medical staff	Family, friends, co-workers	Posters, brochures & others
1. Primary symptoms of dengue infection	44.1	13.2	36.8	60.5	59.9	49.3	29.0	36.8	34.9
2. Incubation period of dengue virus	30.9	6.6	27.0	50.0	44.1	43.4	19.1	15.8	21.7
3. Death rate due to dengue	40.1	11.2	42.1	47.4	43.4	31.6	15.8	12.5	17.8
4. Different treatment options for dengue	21.7	6.6	23.7	60.5	42.8	33.6	42.8	19.1	25.0
5. Availability of dengue-related medicines and vaccines in Singapore	24.3	5.9	28.3	58.6	38.2	30.3	40.1	18.4	23.0
6. Origin of dengue virus	26.3	7.2	26.3	45.4	36.2	53.3	19.1	11.8	13.8
7. Vulnerable groups and their risk levels	38.2	9.9	32.2	52.0	40.8	31.6	25.7	18.4	21.7
8. Potential breeding sites in homes or public areas	46.1	15.8	39.5	45.4	45.4	28.3	15.8	23.0	27.0
9. Peak biting time in a day of dengue mosquitoes	27.6	7.9	24.3	46.7	36.9	36.2	18.4	14.5	19.1
10. Availability of protection products (e.g., insect repellent)	29.6	8.6	28.3	47.4	49.3	27.0	27.0	30.9	19.1
11. Dengue cluster areas in Singapore	44.7	13.2	46.7	50.7	48.0	22.4	9.9	17.1	20.4
12. Current dengue status (e.g., no. of infected cases) in my neighborhood	42.1	14.5	42.8	44.7	44.7	20.4	13.2	23.0	18.4

Values are presented as %.

Table 4. Ranking of top three most preferred communication channels

Information needs	Three top ranked channels		
	1	2	3
1. Primary symptoms of dengue infection	Health websites (hospitals & others)	Social media	General websites, e.g., Wikipedia
2. Incubation period of dengue virus	Health websites (hospitals & others)	Social media	General websites, e.g., Wikipedia
3. Death rate due to dengue	Health websites (hospitals & others)	Social media	Newspapers
4. Different treatment options for dengue	Health websites (hospitals & others)	Social media	Medical staff (doctors, nurses, etc.)
5. Availability of dengue-related medicines and vaccines in Singapore	Health websites (hospitals & others)	Medical staff (doctors, nurses, etc.)	Social media
6. Origin of dengue virus	General websites, e.g. Wikipedia	Health websites (hospitals & others)	Social media
7. Vulnerable groups and their risk levels	Health websites (hospitals & others)	Social media	Television
8. Potential breeding sites in homes or public areas	Television	Health websites (hospitals & others)	Social media
9. Peak biting time in a day of dengue mosquitoes	Health websites (hospitals & others)	Social media	General websites, e.g., Wikipedia
10. Availability of protection products (e.g., insect repellent)	Social media	Health websites (hospitals & others)	Family, friends & co-workers
11. Dengue cluster areas in Singapore	Health websites (hospitals & others)	Social media	Newspapers
12. Current dengue status (e.g., no. of infected cases) in my neighborhood	Newspapers	Health websites (hospitals & others)	Social media

Table 5. Summary of top three preferred channels for seeking dengue-related information (based on 12 information needs)

Ranking	Communication channeled	No. of information needs
1	Websites of hospitals & other health agencies	12 (100)
2	Social media	12 (100)
3	General websites, e.g., Wikipedia	4 (33.3)
4	Newspaper	3 (25.0)
5a	Television	2 (16.6)
5b	Medical staff (e.g. doctors, nurses)	2 (16.6)
7	Family, friends & co-workers	1 (8.3)

Values are presented as number (%).

etc.) were considered suitable for seeking only two out of listed 12 information needs. Basically the respondents were interested in only getting "treatment" related information from medical staff. As medical staff, particularly doctors, are usually busy attending to a large number of patients daily, probably that is why the respondents preferred asking only for essential information from them. For other information needs, they probably thought that they can get this information on their own from other sources and channels.

4.4. Perceptions of Information Quality

Information quality is usually a major concern while seeking health information. The respondents were asked to provide their assessment on five aspects of information quality for different information sources and channels (Table 6). For information accuracy, 113 (74.3%) of the respondents picked medical staff for providing accurate information. Other communication channels considered to provide accurate information by more than one-half of the respondents were: newspapers (63.2%), websites of hospitals and other health agencies (61.2%), and television (59.9%). It was interesting to note that only 11.8% of the respondents felt that information available through social media was accurate although earlier (Table 4) it was among the top three channels for seeking information for all the 12 dengue-related information needs.

It appeared that although a majority of the respondents were using social media for seeking dengue-related information, they were aware that this information may not be completely accurate. It was probably because a majority of the respondents were well-educated and aware of the limitations of information available through social media. However, some senior citizens

Table 6. Perceived quality of dengue information accessible through different channels

Channel		Accuracy	Time-liness	Accessibility	Understand-ability	Adequacy
Television		59.9	44.7	48.7	63.2	34.9
Radio		41.0	37.5	30.3	45.4	23.7
Newspapers		63.2	30.9	50.0	57.2	40.1
Internet	Hospital & health websites	61.2	36.2	59.9	42.1	43.4
	Social media	11.8	49.3	75.0	46.7	21.7
	General websites	18.4	21.1	73.0	48.7	24.3
Medical staff (e.g., doctors, nurses)		74.3	9.9	15.8	39.5	38.8
Family, friends & co-workers		7.2	23.0	58.6	57.9	17.8
Posters, brochures and other materials		45.4	13.8	31.6	59.2	34.9

Values are presented as %.

Table 7. Ranking of communication channels by various information quality attributes

Information attributes	Ranking		
	1	2	3
Accuracy	Medical staff	Health websites (hospitals & others)	Newspapers
Timeliness	Social media	Television	Radio
Accessibility	Social media	General websites	Health websites (hospitals & others)
Understandability	Television	Posters and brochures	Family, friends & co-workers
Adequacy	Health websites (hospitals & others)	Newspapers	Medical staff

and comparatively less educated individuals might not be fully aware of the reliability issue. It is, therefore, desirable that efforts should be made to provide basic information literacy skills to different segments of society (Mokhtar, Majid, & Foo, 2006).

For the time attribute, social media (49.3% of respondents), television (44.7%), and radio (37.5%) were considered as the top three channels for providing timely information. For information accessibility, all the Internet based platforms such as social media (75%), general websites (73%), and websites of hospitals and other health agencies (59.9%) were chosen for easy information accessibility. For information understandability, television (63.2%) was at the top, followed by dengue-related posters and brochures (59.2%), and family members, friends, and co-workers (57.9%). Finally for the information adequacy attribute, the top three channels were websites of hospitals and other health agencies (43.4%), newspapers (40.1%), and medical staff (38.8%).

Table 7 provides ranking of different communication channels based on five information quality attributes.

4.5. General Perceptions of Using Dengue-related Information

A set of 11 statements were used to investigate participants' perceptions of using dengue-related information. A 5-point Likert scale was used for measuring responses, where 1 represented "strongly disagree" and 5 "strongly agree." A majority of the participants agreed (mean score 3.95) that they would more actively look for information if either they or their family members are infected by dengue. A majority of the participants also agreed that usually it is easy to understand dengue-related information (mean score 3.50), the Internet can provide all the necessary information about dengue (mean score 3.25), and that social media is very useful in getting reliable dengue-related information (mean score 3.03) (Table 8).

The remaining 7 statements (i.e., statements 5 to 11) were presented in a negative manner and disagreements with these statements actually represented a positive connotation of the given scenarios. All the negative statements received mean scores of less than 3 which indicate that a majority of the respondents did not agree with these statements. In

Table 8. Respondents' perceptions of using dengue-related information

Statements	Mean score (1–5)	Standard deviation
1. I will actively look for dengue information if I or a family member is infected.	3.95	1.35
2. Usually it is easy for me to understand dengue-related information.	3.50	1.02
3. The Internet can provide all the necessary information about dengue.	3.25	1.02
4. Social media is very useful in getting reliable dengue-related information.	3.03	0.94
5. I don't know how to determine credibility of dengue information available through different information channels.	2.95	1.00
6. Availability of too much dengue information is causing confusion.	2.95	1.00
7. I already know all the necessary information about dengue.	2.81	0.94
8. I feel dengue-related information changes too frequently.	2.71	0.80
9. I receive too much information on dengue from different government agencies.	2.70	0.85
10. Dengue information coming from different agencies is usually contradicting.	2.66	0.91
11. Usually I face difficulty in finding desired dengue-related information.	2.56	0.94

other words, the majority of the participants know how to assess information credibility; too much dengue information was not causing any confusion; they did not think that they know all the necessary information about dengue; dengue-related information was not changing too frequently; too much information was not pushed on them (no information overload); information coming from different agencies was not contradictory; and they were not facing difficulty in getting access to dengue-related information. On the whole, it appeared that the respondents had a good understanding of dengue-related information and were not facing any serious issues in using this information.

5. DISCUSSION AND CONCLUSION

In recent years, the topic of information needs and seeking behavior during epidemics has gained more attention from researchers, and the findings of such studies can be applied in designing effective awareness campaigns. Although for a specific outbreak citizens may have a unique set of information needs, this knowledge can be used for designing awareness campaigns for other similar outbreaks. For example, information needs and seeking behavior for different mosquito-borne diseases such as malaria, dengue, chikungunya, and Zika viruses could have several similarities. This means lessons learnt during one outbreak can be used for developing information communication strategies for possible future epidemics.

This study revealed that the most important dengue-related information needs included information about the availability of dengue-related medicines and vaccines, primary infection symptoms, various available treatments, and dengue cluster areas in Singapore. Previous studies on dengue, H1N1, and H7N9 viruses in different countries had also shown somewhat similar findings. This means that usually the general public is interested in getting information about the symptoms of an epidemic disease, possible treatments, and the necessary preventive measures. Agencies involved in public awareness campaigns can take note of these findings and should try to provide adequate information on these aspects. Currently the Singapore government is actively running dengue awareness campaigns such as "Do the Mozzie Wipeout," mainly focusing on common dengue symptoms, and "Do the 5-step Mozzie Wipeout" with emphasis on key preventive measures. However, it is desirable that the Singapore National Environment Agency should review its campaign messages and try to align these with the information needs of the general public. This review is desirable as a considerable number of respondents said that they do not get access to the necessary dengue-related information. Moreover, the National Environment Agency can also ask its Dengue Prevention Volunteers to share such information during their community visits and public engagements.

Regarding the information seeking behavior, it was found that the participants were heavily dependent on the Internet for getting dengue-related information. This finding is not surprising, as the Internet penetration rate in Singapore is quite

high and according to the Global Information Technology Report, issued by the World Economic Forum (2016), Singapore was ranked first in the world for information technology penetration. According to Singapore InfoComm Media Development Authority, in 2017 around 91% of Singapore households had access to the Internet (InfoComm Media Development Authority, 2018). Thus, heavy dependence on the Internet for seeking dengue-related information is quite understandable. Another possible explanation for high reliance on the Internet could be the demographics of the respondents. A sizeable majority of the respondents were less than 40 years old and possessed a bachelor's degree or a higher qualification. These individuals are likely to be IT literate and comfortable with using the Internet for seeking health as well as dengue-related information.

The most popular source was websites of hospitals and other healthcare agencies. It is understandable as these websites are expected to provide accurate, up-to-date, and comprehensive information about dengue infection. It was also found that social media was also popular, although respondents were aware of the fact that not all the information accessible through this platform is credible. The study revealed that more frequently sought after types of information through social media were: primary symptoms of dengue infection, availability of protective products, and dengue hotspots in Singapore. This indicates that the respondents were familiar with the limitations of social media and were only seeking non-critical dengue-related information through it. They were mostly using authoritative online sources for seeking more critical information, such as different treatment options for dengue infection, and availability of dengue-related medicines and vaccines. The Singapore National Environment Agency should take note of these findings and try to further strengthen its presence on the Web. In addition to revamping its Facebook page, it should also consider other social media platforms such as Twitter and Instagram for disseminating dengue-related and other health information. It is equally important that efforts should be made to enhance information literacy skills of different segments of the society to help them avoid using questionable and low-quality online information.

It is also important to understand that preference for Internet sources does not reduce the power and value of mass media, particularly television and newspapers, for creating awareness about the dengue virus. It is therefore desirable that government agencies should use a variety of communication channels for reaching out to the general public. Although this study was carried out in Singapore, other countries suffering from high rates of dengue infections can also use these findings to decide what information should be disseminated to their citizens and what communication channels will be more appropriate for this purpose.

Overall, the findings of this study provided some useful insights for health information communicators about the types of information required by the general public during health epidemics and their preferred channels for seeking the needed information. In addition, government healthcare agencies, hospitals, public welfare departments, and other agencies involved in designing public health awareness campaigns can use the findings of this study to review their current promotional materials and determine if any changes are desirable. Although the study participants exhibited some knowledge and understanding of the types and quality of information accessible through different sources and communication channels, it is desirable that appropriate measures should be undertaken to further improve the health information literacy of the general public. For this purpose, government agencies can develop online information literacy tutorials and quizzes which can easily be accessed by different segments of the society. Improvement in information literacy skills will empower the general public to seek quality health information effectively and efficiently.

Although the study findings are useful in understanding the information needs and seeking behavior of the general public, it is important to note that these findings cannot be fully generalized to the whole Singapore population. The majority of the study participants were in the age group of 21 to 40 years and possessed a bachelor's degree or a higher qualification. Information needs and seeking behavior of these individuals could be different from teenagers and those individuals who are more than 40 years old. For more comprehensive and generalizable findings, future studies should try to include participants from all age groups with different education levels. Similarly, this study used a survey questionnaire for data collection, whereas future studies can consider using a combination of quantitative and qualitative data collection techniques to collect a more comprehensive dataset.

REFERENCES

Alhazmi, S. A., Khamis, N., Abalkhail, B., Muafaa, S., Alturkstani, A., Turkistani, A. M., & Almahmoudi, S. (2016). Knowledge, attitudes, and practices relating to dengue fever among high school students in Makkah, Saudi Arabia. *International Journal of Medical Science and Public Health*, 5(5), 930-937.

Alves, A. C., Fabbro, A. D., Passos, A. C., Carneiro, A. M., Jorge, T. M., & Martinez, E. Z. (2016). Knowledge and practices related to dengue and its vector: A community-based study from Southeast Brazil. *Revista da Sociedade Brasileira de Medicina Tropical*, 49(2), 222-226.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., . . . Hay, S. I. (2013). The global distribution and burden of dengue. *Nature*, 496(7446), 504-507.

Boonchutima, S., Kachentawa, K., Limpavithayakul, M., & Prachansri, A. (2017). Longitudinal study of Thai people media exposure, knowledge, and behavior on dengue fever prevention and control. *Journal of Infection and Public Health*, 10(6), 836-841.

Brusich, M., Grieco, J., Penney, N., Tisgratog, R., Ritthison, W., Chareonviriyaphap, T., & Achee, N. (2015). Targeting educational campaigns for prevention of malaria and dengue fever: An assessment in Thailand. *Parasites & Vectors*, 8, 43.

Cabrera, R., de la Torre-Del Carpio, A. G., Jesus, A. I. B., Borit, J. M. C., Fuente, F. J. H, Poma. P. V. U., & Ibarra-Casablanca, E. (2016). Knowledge, attitudes and practices about dengue fever in elementary school students in Chorrillos, Lima, Peru [Conocimientos, actitudes y prácticas sobre dengue en estudiantes de educación primaria en Chorrillos, Lima, Perú]. *Anales de la Facultad de Medicina*, 77(2), 129-135.

Chowdhury, P. D., Haque, C. E., & Meyur, S. (2013). Role of primary care providers in dengue prevention and control in the community: Practitioners' and local laypersons' perspectives in Dhaka, Bangladesh. *International Journal of Integrated Care*, 13. Retrieved January 27, 2019 from https://ijic.ubiquitypress.com/articles/10.5334/ijic.1300/galley/2138/download/.

Elsinga, J., Schmidt, M., Lizarazo, E. F., Vincenti-Gonzalez, M. F., Velasco-Salas, Z. I., Arias, L., . . . Tami, A. (2018). Knowledge, attitudes, and preventive practices regarding dengue in Maracay, Venezuela. *American Journal of Tropical Medicine and Hygiene*, 99(1), 195-203.

Firdous, J., Mohamed, A., Al-Amin, M., Ihsan, M., Imadi, M. F., Hakim, M. K., . . . Muhamad, N. (2017). Knowledge, attitude and practice regarding dengue infection among Ipoh community, Malaysia. *Journal of Applied Pharmaceutical Science*, 7(8), 99-103.

Hanim, A. K., Razman, M. R., Jamalludin, A. R., Nasreen, E. H., Phyu, H. M., SweSwe, L., & Hafizah, P. (2017). Knowledge, attitude and practice on dengue among adult population in Felda Sungai Pancing Timur, Kuantan, Pahang. *International Medical Journal Malaysia*, 16(2), 3-9.

Harish, S., Srinivasa, S., Shruthi, P., Devaranavadagi, R. A., Bhavya, G., & Anjum, S. K. (2018). Knowledge, attitude and practice regarding dengue infection among parents of children hospitalized for dengue fever. *Current Pediatric Research*, 22(1), 33-37.

Hassan, S. A., Khail, A. K., Waris, A., Alam, G., & Marwat, S. K. (2017). Assessment of knowledge, attitude and practices regarding dengue fever among adult population of district Dir Lower, Khyber Pakhtunkhwa, Pakistan. *Pakistan Journal of Public Health*, 7(2), 71-74.

InfoComm Media Development Authority. (2018). *Household access to internet 2007-2017*. Retrieved January 27, 2019 from https://www.imda.gov.sg/industry-development/facts-and-figures/infocomm-usage-households-and-individuals#2.

Javed, N., Ghazanfar, H., & Naseem, S. (2018). Knowledge of dengue among students exposed to various awareness campaigns in model schools of Islamabad: A cross-sectional study. *Cureus*, 10(4), e2455.

Lee, K. S., Lo, S., Tan, S. S. Y., Chua, R., Tan, L. K., Xu, H., & Ng, L. C. (2012). Dengue virus surveillance in Singapore reveals high viral diversity through multiple introductions and in situ evolution. *Infection Genetics and Evolution*, 12(1), 77-85.

Li, T., Feng, J., Qing, P., Fan, X., Liu, W., Li, M., & Wang, M. (2014). Attitudes, practices and information needs regarding novel influenza A (H7N9) among employees of food production and operation in Guangzhou, Southern China: A cross-sectional study. *BMC Infectious Diseases*, 14(4), 1-12.

Majid, S., & Rahmat, N. A. (2013). Information needs and seeking behavior during the H1N1 virus outbreak. *Journal of Information Science Theory and Practice*, 1(1), 42-53.

Mokhtar, I. A., Majid, S., & Foo, S. (2006). Teaching information literacy through multiple intelligences and mediated learning: A quasi-experimental study. *Singapore Journal of Library and Information Management*, 35(3), 10-25.

National Environment Agency. (2018). *Dengue: Quarterly dengue surveillance data*. Retrieved January 27, 2019 from https://www.nea.gov.sg/dengue-zika/dengue/quarterly-dengue-surveillance-data.

Nivedita. (2016). Knowledge, attitude, behaviour and practices (KABP) of the community and resultant IEC leading to behaviour change about dengue in Jodhpur City, Rajasthan. *Journal of Vector Borne Diseases*, 53(3), 279-282.

Ooi, E. E., Goh, K. T., & Gubler, D. J. (2006). Dengue prevention and 35 years of vector control in Singapore. *Emerging Infectious Diseases*, 12(6), 887-893.

Rajarethinam, J., Ang, L.-W., Ong, J., Ycasas. J., Hapuarachchi, H. C., Yap, G., . . . Ng, L.-C. (2018). Dengue in Singapore from 2004 to 2016: Cyclical epidemic patterns dominated by serotypes 1 and 2. *American Journal of Tropical Medicine and Hygiene*, 99(1), 204-210.

Seltenrich, N. (2016). Singapore success: New model helps forecast dengue outbreaks. *Environmental Health Perspectives*, 124(9), A167.

Siddiqui, T. R., Ghazal, S., Bibi, S., Ahmed, W., & Sajjad, S. F. (2016). Use of the health belief model for the assessment of public knowledge and household preventive practices in Karachi, Pakistan, a dengue-endemic city. *PLoS Neglected Tropical Diseases*, 10(11), e0005129.

Singapore Ministry of Health. (2018). *Dengue fever in Singapore*. Retrieved January 27, 2019 from https://www.healthhub.sg/a-z/diseases-and-conditions/192/topic_dengue_fever_MOH.

Siriwardana, E., & Samarasinghe, K. (2018). Secondary school teachers' knowledge, attitudes and preventive practices of dengue fever. *GSTF Journal of Nursing and Health care*, 5(1).

Viennet, E., Ritchie, S. A., Williams, C. R., Faddy, H. M., & Harley, D. (2016). Public health responses to and challenges for the control of dengue transmission in high-income countries: Four case studies. *PLoS Neglected Tropical Dieases*, 10(9), e0004943.

Voeten, H. A., de Zwart, O., Veldhuijzen, I. K., Yuen, C., Jiang, X., Elam, G., . . . Brug, J. (2009). Sources of information and health beliefs related to SARS and avian influenza among Chinese communities in the United Kingdom and the Netherlands, compared to the general population in these countries. *International Journal of Behavioral Medicine*, 16(1), 49-57.

Wong, L. P., & Sam, I. C. (2010). Public sources of information and information needs for pandemic influenza A (H1N1). *Journal of Community Health*, 35(6), 676-682.

World Economic Forum. (2016). *The global information technology report 2016: Innovating in the digital economy*. Retrieved January 27, 2019 from http://www3.weforum.org/docs/GITR2016/WEF_GITR_Full_Report.pdf.

World Health Organization. (2019a). *What is dengue?* Retrieved January 27, 2019 from http://www.who.int/denguecontrol/disease/en/.

World Health Organization. (2019b). *Dengue control: Epidemiology*. Retrieved January 27, 2019 from https://www.who.int/denguecontrol/epidemiology/en/.

Yboa, B. C., & Labrague, L. J. (2013). Dengue knowledge and preventive practices among rural residents in Samar province, Philippines. *American Journal of Public Health Research*, 1(2), 47-52.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 29-38, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.3

Comparison of User-generated Tags with Subject Descriptors, Author Keywords, and Title Terms of Scholarly Journal Articles: A Case Study of Marine Science

Praveenkumar Vaidya*
Department of Studies in Library and Information Science,
University of Mysore, Mysuru, India
Tolani Maritime Institute, Pune, India
E-mail: praveenv@tmi.tolani.edu

N. S. Harinarayana
Department of Studies in Library and Information Science,
University of Mysore, Mysuru, India
E-mail: ns.harinarayana@gmail.com

ABSTRACT
Information retrieval is the challenge of the Web 2.0 world. The experiment of knowledge organisation in the context of abundant information available from various sources proves a major hurdle in obtaining information retrieval with greater precision and recall. The fast-changing landscape of information organisation through social networking sites at a personal level creates a world of opportunities for data scientists and also library professionals to assimilate the social data with expert created data. Thus, folksonomies or social tags play a vital role in information organisation and retrieval. The comparison of these user-created tags with expert-created index terms, author keywords and title words, will throw light on the differentiation between these sets of data. Such comparative studies show revelation of a new set of terms to enhance subject access and reflect the extent of similarity between user-generated tags and other set of terms. The CiteULike tags extracted from 5,150 scholarly journal articles in marine science were compared with corresponding Aquatic Science and Fisheries Abstracts descriptors, author keywords, and title terms. The Jaccard similarity coefficient method was employed to compare the social tags with the above mentioned wordsets, and results proved the presence of user-generated keywords in Aquatic Science and Fisheries Abstracts descriptors, author keywords, and title words. While using information retrieval techniques like stemmer and lemmatization, the results were found to enhance keywords to subject access.

Keywords: Web 2.0, social tagging, information retrieval, Jaccard similarity, subject descriptors

Open Access

Accepted date: February 28, 2019
Received date: July 27, 2018

***Corresponding Author:** Praveenkumar Vaidya
Librarian
Tolani Maritime Institute, Induri, Talegaon, Pune 410507, India
praveenv@tmi.tolani.edu

1. INTRODUCTION

Information retrieval in the context of information overload is the challenge for library and information architects. The adversity in recalling relevant information with precision is exacerbated when substantial information afforded by the Internet is available in abundance. In order to organize such profusely accessible information, library professionals have designed many hierarchical classification systems or subject related controlled vocabularies. The shift in this order arose due to the impact of advancement in Web 2.0 (Anfinnsen, Ghinea, & de Cesare, 2011) applications wherein many social networking platforms enabled users to organize their personal information resources in the form of social tags or folksonomies. Hence, the folksonomies are user created metadata (Furner, 2010; Guy & Tonkin, 2006; Wal, 2004) for web resources and are used extensively for content categorization and retrieval in the age of Web 2.0. Unlike a controlled vocabulary which is designed by top-down ways, a folksonomy is constructed from bottom-up by user-centred ways to organize personal information resources.

Mathes (2004) indicates about three groups which are predominantly involved in providing keywords to resources which are also used for effective retrieval: the authors, users, and subject experts. But generally, the keywords provided by subject experts, known as controlled vocabulary, are a popular dataset. The hierarchical structure of subject-specific taxonomies is prevalent in knowledge organisation but with some limitation (Golder & Huberman, 2006; Kipp, 2006). In the case of author-assigned keywords, authors are normally asked to choose a few keywords which describe the content of their own article (Névéol, Doğan, & Lu, 2010), but which may not be sufficient to greater precision and recall. Furthermore, user-generated keywords or collaborative tags have the ability to facilitate both retrieval and discovery. Folksonomies can be navigated through tags, resources, and users for any user query within a single user-centric environment for effective retrieval system. Hence, tags can also be a useful dataset for content categorization and knowledge organisation (Peters et al., 2011; Rafferty, 2017; Stan & Maret, 2017). In scholarly journal articles or any other source of the document the 'title'grabs the attention of the user at first sight. Therefore for any researcher the 'title'plays an important role that provides aboutness and contents of the document. Hence, the title terms are also a dominant source of metadata in information retrieval (Davarpanah & Iranshahi, 2005; Voorbij, 1998).

As 'social tags'represent a tagger's conceptual understanding or categorization of a resource from a personal point of view, hence researchers consider social tagging as related to sense making (Hotho, Jäschke, Schmitz, & Stumme, 2006). The 'subject descriptors'or index terms, which are also descriptive metadata like social tags, come from highly structured controlled vocabularies. The 'author keywords'consist of conceptual and content categorization from the author's perspective, and add important value to resources. Similarly, title words accurately describe the contents of the manuscript, hence are presented as significant metadata. Given their conceptually shared purpose of social tags, subject taxonomies, author keywords, and title words, it makes sense to investigate whether social tags can complement subject descriptors, author keywords, and title words. Essentially, the purpose of this work is to understand whether social tags can also emerge as alternative access points to subject access despite the presence of subject descriptors, author keywords, and title words.

All these above-mentioned datasets have some limitations in precise retrieval and hence need to be studied for useful application. The combination of folksonomy, controlled vocabulary systems, author-supplied keywords, and title terms is an effective way to make up for the shortcomings of all these metadata for effective information retrieval.

2. LITERATURE REVIEW

There are many studies where comparative works are done to understand the significance of the datasets. Such studies demonstrated the emergence of additional useful terms for search and information retrieval which also enhance the process of knowledge discovery.

Several studies are found where comparison of datasets is conducted between social tags and subject descriptors. In their study, C. Lu, Park, and Hu (2010) examined the "difference and connections between social tags and expert-assigned subject terms and further explored the feasibility and obstacles of implementing social tagging in library systems. The results show the possible use of social tags to improve the accessibility of library collections."In another study, Wu, He, Qiu, Lin, and Liu (2013) believe that tagging has the potential to become a complementary resource for expanding and enriching controlled vocabulary systems. They also propose that "the help of future technology to regulate and promote features related to controlled vocabulary in social tags would greatly improve people's organizational and access capabilities within information resources."Hence, there was an attempt to enhance information retrieval using social tags in addition to subject vocabularies.

But this comparison work also involves author keywords and title words in addition to subject descriptors to compare them with social tags. In one of the early studies on comparing user, creator, and intermediary tagging, Kipp (2006) examined these three set of words and found the presence of many user terms which were related to the author and controlled vocabularies. A few terms were also found which were not available in controlled vocabularies and it was concluded that user tags can provide additional access points to discover information. In other studies by Kipp (2011a, 2011b), similar datasets were compared and analysed by using descriptive statistics method, informetric measures, and thesaural term comparison. The results showed the presence of additional access terms in tags and it was recommended to take advantage of these terms over traditional systems.

Similarly, Lu and Kipp (2014) and Syn and Spring (2010) conducted studies to evaluate whether user tags can represent resources as author keywords do and are used to categorize resources as keywords. The cosine similarity test was conducted to measure the similarity value. The results showed that author provided keywords were more consistent in describing the content of the resources. But, the user-assigned tags showed more variation in describing the content of the resources. In the same study, the researchers also conducted a comparative study of both the title and abstract terms of papers. In case of comparison of tags with title keywords, it was observed that the title of papers seems to be the main source for users to assign tags and therefore tags and title keywords represent the content of the paper.

In another early study Voorbij (1998) compared title keywords with subject descriptors to demonstrate that the subject descriptors retrieve more precise and far more successful results than by searching through title keywords. The study concludes that many relevant records cannot be retrieved by title keywords because of the wide diversity of ways to express the topic. In a similar study, Ansari (2005) tried matching between assigned descriptors and title keywords of medical theses. The results show that the keywords in the title comprise genuine information value and it was recommended that such words should be taken into consideration while introducing them into the indexing descriptors. The other study by Engelson (2013) worked to determine the correlation between title keywords and Library of Congress Subject Headings (LCSH) terms, and found that books with a popular content level designator had high-level matches.

Strader (2009) examined the overlap between author-assigned keywords with LCSH terms. It was observed that both keywords and controlled vocabularies complement one another and the ability to provide unique access points for the majority of the searches was demonstrated. But both LCSH and keywords provide significant numbers of unique terms that may increase the discoverability of resources.

The above studies suggest that the comparison of user assigned tags with author keywords, title words, and subject descriptors will result in new access points to information discovery and retrieval.

This study stands apart due to comparison work undertaken with different datasets and methodology as well. The CiteULike tags have been compared with subject descriptors, author keywords, and title words also. Even though the above review shows such works, they differ in the methodology adopted for this work. In some other works, where the same methodology is adopted, they differ in the datasets considered for this work.

3. RESEARCH QUESTIONS

In this study, an attempt is made to address the following research questions:

A. Is there any similarity between CiteULike tags with Aquatic Science and Fisheries Abstracts (ASFA) subject descriptors, author-assigned keywords, and title terms of marine science literature?

B. Do social tags enhance the effectiveness of keywords to subject access better than controlled descriptive terms, author-assigned keywords, and title terms?

The findings of this research work will exhibit the importance of social tags for information retrieval and knowledge organisation.

4. SCOPE AND METHODOLOGY OF THE STUDY

Essentially, for this research work the user-generated tags were primary data which were extracted from the social bookmarking site CiteULike. CiteULike is a popular social web service where users can save and share citations from scholarly journal articles. With its great compatibility with subject databases and publishers, it can capture bibliographic data of research articles. This also provides an opportunity to users to annotate personal keywords (tags) to the articles for repeat access. Not only are these tags personally useful, but also are to other researchers of the same field. If a profile is created with subject interests, users can join them and idea

exchange can be facilitated to access the reference articles of other researchers at one place and understand the research carried out by peers. CiteULike also allows users to import/export the citation details in many formats. The tags created by many such users can be useful for research work. As CiteULike is popular among researchers it attracts listings of many articles and a good number of social tags also. Hence, CiteULike has an edge over other available reference management tools. For this research work, marine science scholarly journal articles were chosen due to the dynamic nature of the subject. Globally, between 2010 and 2014 more than 370,000 manuscripts were published and more than 2 million articles were cited in marine science. The research and development expenditure of countries with high gross domestic product show high ocean science performance in terms of publications and citations (United Nations Educational, Scientific and Cultural Organization, 2017).

Marine science journal titles were collected from the list of ASFA. Consequently, the researcher identified and gathered 5,150 articles from the ASFA journal list published during 1954 to 2015, in which 1,405 articles belonged to publication year 1954 to 2000 and the remaining 3,745 were published during 2001 to 2015. The collected journal articles were searched in CiteULike to collect the tags, which resulted in 42,369 tags from 356 marine science journals. Similarly, these articles were also searched in the ASFA database to collect the corresponding subject headings and author keywords. WebCorp, an online tool, was used to convert the selected titles into a wordlist. All these datasets were transposed to Excel (Microsoft, Redmond, WA, USA) to manipulate the data. For these 5,150 articles the corresponding 49,478 subject headings, 10,752 author-assigned keywords, and 8,019 title terms were accumulated (Table 1). The research did require unique words, hence all these datasets were tested for duplication and the overlapped words were removed. For stemmer and lemmatization, it was also necessary to convert datasets to single words. Hence, during this process of converting multi-words to single words, the researcher could uncover 6,391 unique CiteULike tags, 5,695 ASFA words, 6,391 author keywords, and 7,213 title words.

The extracted CiteULike tags were transposed to Microsoft Excel sheets and the tags were preprocessed by removing the trashy tags (Thomas, Caudle, & Schmitz, 2010). These chosen articles were searched in the ASFA database and the corresponding controlled vocabularies were collected which were also transposed to Microsoft Excel sheets. Simultaneously, these 5,150 articles were explored with their DOI and author keywords were mined. Additionally, from these selected scholarly articles, title terms were also created. The user-generated CiteULike tags were compared with ASFA controlled vocabularies, author keywords, and title words to recognize to what extent these user tags resemble and enhance the keywords to subject access.

Table 1. Summary of all distinct dataset of each type

Datasets	Total words	Distinct words
CiteULike tags	42,369	9,015
ASFA descriptors	49,478	10,106
Title words	8,019	8,019
Author keywords	10,752	6,545

4.1. Preprocessing of Words

Preprocessing of CiteULike tags, author keywords, and title words is a very significant process for effective comparison work. Generally, the social tags were likely to consist of unpredictable and inconsistent words assigned by different users, unlike the controlled vocabularies which are organised and hierarchical in nature. The tags were assigned with several variations of singular, plural, hyphen, underscore, words with numerals or just numerals, acronyms/abbreviations, compound words, and also foreign language words. The tags have a serious deficit of synonym control and lack of precision and recall, which creates a challenge for retrieval effectiveness. Such inconsistencies are natural because the social tags are user-oriented, collaborative, democratic, cheap, dynamic, and distributed. CiteULike prevents users from assigning two words for any resource. Hence, users assign tags interpolated with 'underscore'or 'hyphen'between two concepts. With such limitations, these CiteULike tags must be normalized to compare them with ASFA descriptors, author keywords, and title words. Furthermore, these tags were converted into more meaningful words by removing hyphens, underscores, numerals, and foreign words.

This research work also includes the process of stemming and lemmatization of CiteULike tags, ASFA descriptors, author keywords, and title words. Hence, these datasets were converted to single words. Such single words were stemmed/lemmatized (http://text-processing.com/demo/stem/) to reduce variants (Mohammad, 2018). The stemmer is a function used in many text retrieval systems and search engines. A stemming algorithm is used to reduce words to their stems or roots after removing their prefixes and suffixes. For example, 'activatio n,''active,''activities,'and 'activity'were stemmed to 'activ.'With stemmer, these four words will turn into a single word, which enhances the efficiency of comparison (Lee & Schleyer, 2010,

2012; Syn & Spring, 2013). On the other hand, lemmatization works on morphological analysis of words and tries to remove inflectional endings thereby returning words to their dictionary form. For example, the word form of 'studies'and 'studying'was lemmatized to 'study'in both cases (Risueño, 2018).

4.2. Comparison of Words by Jaccard Similarity Coefficient

A similarity coefficient represents the similarity between two set of keywords, two documents, two queries, or one document and one query. A similarity coefficient is a function which computes the degree of similarity between a pair of text objects (Heymann & Garcia-Molina, 2009; Niwattanakul, Singthongchai, Naenudorn, & Wanapu, 2013; Thada & Jaglan, 2013).

The Jaccard Similarity coefficient is used to measure the similarity between the frequent sets of tags and the terms employed in the dataset (C. Lu et al., 2010). The Jaccard similarity index is a statistical tool used to compare similarity and diversity of sample datasets and is defined as the size of the intersection divided by the size of the union of the sample dataset. For example, A is the tag dataset comprising distinct tags for articles and B is the term dataset, comprising distinct terms for articles. The Jaccard similarity, ranging from 0 to 1, suggests the amount of overlap between the two data sets (C. Lu, Zhang, & He, 2016). This can be represented in the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

(if A and B are both empty, we define J (A, B) = 1 where $0 \leq J(A,B) \leq 1$)

4.3. Comparison of CiteULike Tags with ASFA Descriptors, Author Keywords, and Title Words

In this work, the CiteULike tags were compared with

- ASFA descriptors
- author keywords and
- title words

Furthermore, the comparison is tested in four different formats of following word structures. All CiteULike tags were compared with ASFA descriptors, author keywords, and title words in the following manner:

- Preprocessed CiteULike tags were compared with preprocessed ASFA descriptors, preprocessed author keywords, and preprocessed title words.
- Lemmatized CiteULike tags were compared with lemmatized ASFA descriptors, lemmatized author keywords, and lemmatized title words.
- Stemmed CiteULike tags were compared with ASFA stemmed descriptors, stemmed author keywords, and stemmed title words.
- CiteULike single tags were compared with ASFA single descriptors, single author keywords, and single title words.

The similarity or overlap between these sets of words was measured by adopting Jaccard's coefficient method, which is the commonly used method in such studies that also helps to answer the research questions considered for this study.

5. ANALYSIS AND INTERPRETATION

It is interesting to know the results of the comparison study conducted for this work. It will be tested to what extent the CiteULike tags will overlap with subject descriptors, author keywords, and title words. This comparison is tabulated in the form of tables for the benefit of understanding in detail. For this comparison work, Microsoft Excel functions were used in an extensive manner. As the extracted data runs into thousands of rows, Microsoft Excel was used for data manipulation. The following tables reveal the outcome and analysis of comparative study between different datasets. The results showed that Jaccard similarity was enhanced when the CiteULike tags, ASFA descriptors, author keywords, and title words were employed with stemmer, lemmatizer, and single words. But information retrieval depends on precision and recall. It was interesting to notice that when the terms were stemmed and lemmatized the Jaccard similarity index was consistently improved more than the results of preprocessed and single words to indicate the presence of more common terms in the datasets. It was also observed that the non-similar terms also play a vital role in such comparative studies.

5.1. Comparison Between CiteULike Tags with ASFA Descriptors

Table 2 shows a glimpse of comparison of terms between CiteULike tags with ASFA descriptors. Table2 illustrates the similarity measurement between CiteULike tags and ASFA descriptors. The results show that the Jaccard coefficient is just 9.17% when compared with preprocessed words of both datasets, which is minimal in the context of parameters of comparison. However, the results show maximum similarity when these words were stemmed (30.73%). But it was also observed that when these words were either lemmatized or

stemmed the similarity seems to go higher and they are in close proximity to each other (25.87% and 30.73%), but the similarity is reduced to 22.73% when compared with single tags of CiteULike and ASFA descriptors, which indicates the importance of stemmer or lemmatized words or even single words for retrieval. Further, whenever there is a high rate of similarity the effectiveness of retrieval also increases, but may hamper precision.

It was also observed that the Jaccard index, when compared with preprocessed CiteULike tags and ASFA descriptors, is 9.17% whereas when the tags and descriptors are converted to single words, the Jaccard similarity index rose to 22.73%. This describes the importance of splitting tags or descriptors into single words to find the common words in order to enhance retrieval efficiency. Due to the splitting of words, the retrieval precision may be affected but recall will be enhanced.

Additionally, this comparison between CiteULike tags and ASFA shows that users do not really have any knowledge of subject taxonomies. The tags were assigned to the sources which were convenient for users to recall and retrieve when needed. Due to this, there may be just 9.17% of common words or similar words, which is very low. However, controlled vocabularies play a vital role for precise information retrieval and hence cannot be neglected (Heymann & Garcia-Molina, 2009; Lee & Schleyer, 2010, 2012; C. Lu et al., 2010; C. Lu et al., 2016; Wu et al., 2013). It was also noticed that some important words which were present in tags but did not find a place in taxonomies may help to enhance the taxonomical dataset, which in turn may help in retrieval precision. For example, 'accretionary wedge'was listed in tags but did not find a place in ASFA descriptors. Similarly, the term 'nutrient starvation'was recorded in user tags, but in ASFA it was registered as 'nutrient deficiency'and 'nutrient depletion.'Therefore, the terms available in tags can also be used as 'Related Term'in controlled vocabulary entries.

5.2. Comparison of CiteULike Tags with Author Keywords

Author keywords are an integral part of the articles and these keywords were compared with user-generated CiteULike tags. Table 3 demonstrates the comparison of CiteULike tags with author keywords. As mentioned earlier, author keywords characterize the content of the research work published in any document. The author keywords are always considered as an important feature of information retrieval.

In this case, the CiteULike tags were compared with author keywords and interesting results were found. It was witnessed that when preprocessed CiteULike tags were compared with author keywords, the overlap was relatively higher or almost double (19.21%) than for the ASFA descriptors, as suggested in Table 2 (9.17%). It infers that users were probably influenced by keywords provided by authors. Hence the overlap was 19.21% between CiteULike tags and author keywords.

And it was also noticed that when these tags and keywords were converted into single words, the Jaccard index of the overlap was 39.61%, which is considerably high. Subsequently, when these same words were stemmed the overlap rose to 44.47%. Besides this, even when compared with lemmatized, stemmed, and single words, the overlap results show high in the case of stemmed words. It can also be understood that when the words are stemmed the overlapped result was the highest (44.47%) among these three entities. Conversely, it was also true that the author keywords and social tags did not match to a large extent (80.79%) and differ in the context of assigning. Similarly, the comparison with CiteULike tags and author keywords also indicated that the author keywords were not matched by around 59.44% when compared as single words. This also indicates that there is a distinct difference between the context of the user and author of the article (Kipp, 2006, 2007). Both author and user think in a diverse direction while assigning keywords to the article.

Table 2. Comparison of CiteULike tags with ASFA words

Datasets	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
CiteULike tags	9,015	6,391	6,391	6,391
ASFA words	10,106	5,695	5,695	5,695
Common words	1,606	2,484	2,841	2,238
Jaccard index	0.0917 (9.17%)	0.2587 (25.87%)	0.3073 (30.73%)	0.2273 (22.73%)

ASFA, Aquatic Science and Fisheries Abstracts.

Table 3. Comparison of CiteULike tags with author keywords

Datasets	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
CiteULike tags	9,015	6,391	6,391	6,391
Author keywords	6,545	4,261	4,261	4,261
Common words	2,507	3,022	3,279	3,074
Jaccard index	0.1921 (19.21%)	0.3961 (39.61%)	0.4447 (44.47%)	0.4056 (40.56%)

As discussed above, more similarity in tags indicates that users also tend to derive the tags from the author keywords. This can be seen by looking into long multi-word keywords. Author keywords like 'altricial versus precocial development,''taxonomic and functional approaches,'and 'western and central pacific fisheries commission'appeared in the dataset of CiteULike tags. These author keywords were mentioned as tags by the users.

5.3. Comparison of CiteULike Tags with Title Terms

Title words play an important role in information retrieval as controlled vocabularies and author keywords. In this section, CiteULike tags were compared with title terms and analysed for their overlap in the context of social tags, as title words are also one of the important datasets for retrieval. Table 4 explains the comparison of these two datasets and analysis is explained for better understanding.

By observing Table 4, there is 38.93% of overlap with CiteULike tags when these words were lemmatized, while in single words the overlap is 36.29%. In Tables 2 and 3, the rate of overlap was more in the case of stemmed words (30.73% and 44.47%), while in Table 4 the similarity result shows 22.7%, which is quite less than in Table 2 and 3. However, it can be noted that social tags were derived from both title terms (36.29%) and author keywords (40.56%) significantly. The reverse is also true for social tags where users not only rely on author and title keywords but they also prefer to provide tags, whichever was convenient for them and for their personal retrieval.

Table 4 also indicates the common terms between CiteULike tags and title terms. The presence of common terms was 2,986 when comparison was done with preprocessed words, which signifies the user was influenced by the title of the article to assign tags. The similar terms were more when CiteULike tags were compared with ASFA descriptors and author keywords. When lemmatized tags and title words were compared the Jaccard ratio was found to be 38.93%, which was more than the comparative result of stemmed words (22.7%) and also preprocessed words (21.26%). Hence it can be also derived that the processed words yield poor Jaccard values, in comparison with lemmatized, stemmed, or single words.

5.4. Comparison of the Jaccard Index of All Datasets

It is essential to analyse the Jaccard index of all these compared datasets. With reference to Table 5, the Jaccard index of all these datasets suggests that the tags and keywords throw consistent results when they were either lemmatized or stemmed or in the form of single words. While the words or tags before preprocessing narrowly result in any significant overlaps. Hence there is a need to determine the retrieval richness, if words were in single, lemmatized, or stemmed format.

This Jaccard index of the words before preprocessing indicates a very low overlap for datasets produced by users, experts, and authors because the titles to scholarly articles were also provided by authors. This does mean that the terms assigned by users, experts, and authors were very different, and even though a few terms are very popular among users, they are not used by experts and vice versa (C. Lu et al., 2010).

Table 5 shows that social tags were in higher agreement with author keywords and title words while describing the content with the controlled vocabularies. The Jaccard index for author keywords (19.21%) and title words (21.26%) was almost the same compared to controlled vocabularies (9.17%). This indicates less overlap in controlled vocabularies in respect to comparison with author keywords and title words. Fig. 1 provides a graphical representation of the same.

With the usage of more techniques like lemmatization and stemmer the researcher had tried to reduce the lexical variation and compared them to enhance the overlap which is visible from the above tables and figure. The Jaccard index for lemmatized words, stemmed words, and single terms stands higher than preprocessed words. This indicates that after preprocessing the tags the rate of similarity or overlap will increase considerably and provide rich dividends in retrieval but may also affect precision.

Table 4. Comparison of CiteULike tags with title words

Datasets	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
CiteULike tags	9,015	6,391	6,391	6,391
Title words	8,019	7,213	7,213	7,213
Common words	2,986	3,812	2,517	3,622
Jaccard index	0.2126 (21.26%)	0.3893 (38.93%)	0.2270 (22.70%)	0.3629 (36.29%)

Table 5. Jaccard index of compared dataset

Comparison of CiteULike tags with	Words before preprocess	Lemmatized words	Porter stemmer words	Single words
ASFA words	0.0917	0.2587	0.3073	0.2273
Author keywords	0.1921	0.3961	0.4447	0.4056
Title words	0.2126	0.3893	0.2270	0.3629

ASFA, Aquatic Science and Fisheries Abstracts.

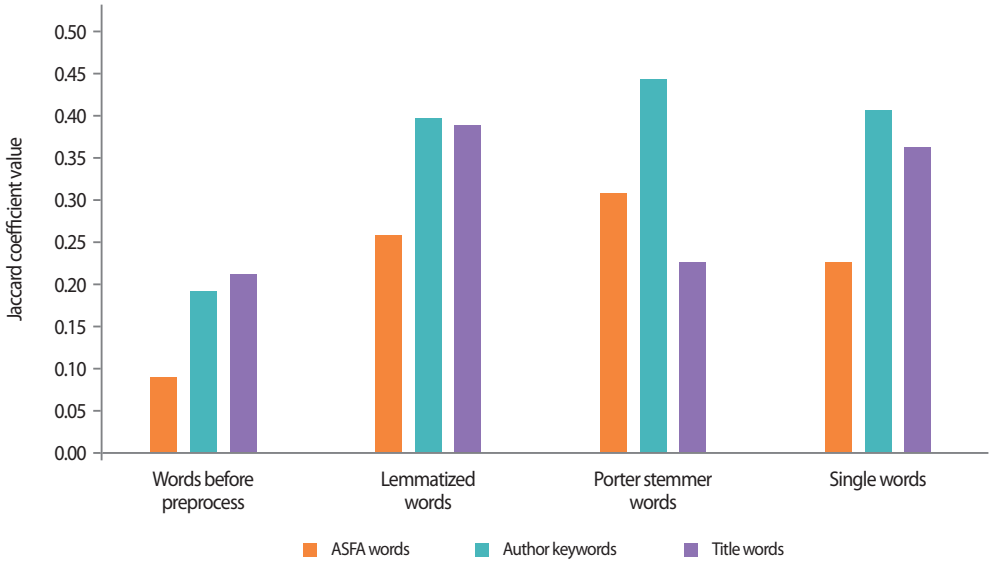


Fig. 1. Jaccard index of comparison of CiteULike tags with other datasets. ASFA, Aquatic Science and Fisheries Abstracts.

6. DISCUSSION AND CONCLUSION

The emergence of Web 2.0 technology has provided an enormous opportunity for users to access their resources by assigning tags to their sources, and typically 'social tagging'allows users to participate and interact with professionals. However, social tags have limitations because of their more uncontrolled and inconsistent nature, and scepticism exists about the value of these tags. In this context, this research work tried to address some of the reservations of social tags and an attempt has been made to throw light on the effectiveness of social tags for the retrieval process and in what way the tags may enhance subject access.

In this research work, the researchers have presented an exhaustive assessment of the association between ASFA (controlled vocabularies), author keywords, and title terms in comparison with CiteULike tags, particularly in the domain of marine science by using Jaccard similarity coefficient method. In context to the research questions of this study, the comparison task was conducted between ASFA descriptors with CiteULike tags (Table 2) to determine the presence of similar or overlap words among them. The result shows a minimal existence (9.17%) of similar words was found when compared before preprocessing. However, the similarity compliance was enhanced when these words were subjected to text processing techniques like lemmatization and stemmer to reduce lexical variation. As a result, the similarities were increased up to 30.73%. These results adequately answer the research question A, which emphasize the presence of common or overlap terms between these datasets. Hence, the users and experts share common terms even though lexical overlap between the corpora is very negligible.

The author keywords and title terms were considered to indicate the content of the article published and their respective CiteULike tags may comprise of tags significantly related to the article. The comparison work showed also in Table 3 and 4 that suggested the rise in overlap or similarity (19.21% and 21.26%) against CiteULike tags.

In an attempt of comparison between CiteULike tags and ASFA descriptors, author keywords, and title words, it was implied that the user was mostly influenced by either author keywords or title words before assigning the tags to resources. The comparison of tags with author keywords and title words resulted in good similarity ratios, which attributes the importance of author keywords and title words. This work clearly illustrated the gain in retrieval when the datasets were compared in single, lemmatized, or stemmed format. The implication of this study can be summarized that information retrieval can be enhanced when multiple words were split into single words but this may affect the precision. Future study could involve finding the appropriate techniques for precision retrieval.

However, it is interesting to know whether social tags can enhance 'subject access'in comparison with the subject

descriptors, author keywords, or title words. This comparison work emphasizes that there is an overlap of terms among subject descriptors, author keywords, and title words. But it can be very well presumed that the non-overlap tags also convey 'subject access'value in the ASFA database, as these CiteULike tags are subject specific to marine science. For example, when the preprocessed CiteULike tags in the form of single words were compared with ASFA single terms the Jaccard index was found to be 22.73%. The non-overlapped terms in the dataset, which is also known as 'Jaccard distance,'was found to be 77.27%. This can be further elaborated, as the presence of 4,153 non-overlapping terms in the dataset has also 'subject access'value. These words may be absent in ASFA yet may throw search results related to the subject, but may hamper precision. This explanation reflects research objective B considered for this study, which specifies the enhancement of keywords to subject access.

Overall, the introduction of social tags in the Web 2.0 context is an opportunity for libraries to enhance their access to resources. Many studies have concluded that their overlap is relatively low but still are very different in their nature and cannot be neglected, and also similarly cannot be considered as an alternative schema for a controlled vocabularies system. However, the user generated tags have a potential to become a complementary source to enhance and enrich a controlled vocabulary system which has the presence of multiple semantic relationships between them.

With the help of semantic technology the integration of social tags and controlled vocabularies can be achieved. With this combination there is a possibility to improve the access and organisation of information resources.

REFERENCES

Anfinnsen, S., Ghinea, G., & de Cesare, S. (2011). Web 2.0 and folksonomies in a library context. *International Journal of Information Management*, 31(1), 63-70.

Ansari, M. (2005). Matching between assigned descriptors and title keywords in medical theses. *Library Review*, 54(7), 410-414.

Davarpanah, M. R., & Iranshahi, M. (2005). A comparison of assigned descriptors and title keywords of dissertations in the Iranian dissertation database. *Library Review*, 54(6), 375-384.

Engelson, L. (2013). Correlations between title keywords and LCSH terms and their implication for fast-track cataloging. *Cataloging & Classification Quarterly*, 51(6), 697-727.

Furner, J. (2010). Folksonomies. In M. J. Bates, & M. N. Maack (Eds.), *Encyclopedia of library and information sciences* (3rd ed.). New York: Taylor and Francis.

Golder, S. A, & Huberman, B. A. (2006). Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32(2), 198-208.

Guy, M., & Tonkin, E. (2006). Folksonomies: Tidying up tags? *D-Lib Magazine*, 12(1). Retrieved July 30, 2018 from http://www.dlib.org/dlib/january06/guy/01guy.html.

Heymann, P., & Garcia-Molina, H. (2009). *Contrasting controlled vocabulary and tagging: Do experts choose the right names to label the wrong things?* Paper presented at the Second ACM International Conference on Web Search and Data Mining (WSDM), Barcelona, Spain.

Hotho, A., Jäschke, R., Schmitz, C., & Stumme, G. (2006). Information retrieval in folksonomies: Search and ranking. In Y. Sure & J. Domingue (Eds.), *The semantic web: Research and applications* (pp. 411-426). Berlin/Heidelberg: Springer.

Kipp, M. E. I. (2006). *Exploring the context of user, creator and intermediary tagging*. Retrieved July 30, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.172.9783&rep=rep1&type=pdf.

Kipp, M. E. I. (2007). *Tagging practices on research oriented social bookmarking sites*. Retrieved July 30, 2018 from http://hdl.handle.net/10150/105837.

Kipp, M. E. I. (2011a). Tagging of biomedical articles on CiteULike: A comparison of user, author and professional indexing. *Knowledge Organization*, 38(3), 245-261.

Kipp, M. E. I. (2011b). User, author and professional indexing in context: An exploration of tagging practices on CiteULike. *Canadian Journal of Library and Information Science*, 35(1), 17-48.

Lee, D. H., & Schleyer, T. (2010). *A comparison of meSH terms and CiteULike social tags as metadata for the same items*. Paper presented at the 1st ACM International Health Informatics Symposium, Arlington, VA, USA.

Lee, D.H., & Schleyer, T. (2012). Social tagging is no substitute for controlled indexing: A comparison of Medical Subject Headings and CiteULike tags assigned to 231,388 papers. *Journal of the American Society for Information Science and Technology*, 63(9), 1747-1757.

Lu, C., Park, J., & Hu, X. (2010). User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings. *Journal of Information Science*, 36(6), 763-779.

Lu, C., Zhang, C., & He, D. (2016). Comparative analysis of book tags: A cross-lingual perspective. *The Electronic*

Library, 34(4), 666-682.

Lu, K., & Kipp, M. E. I. (2014). Understanding the retrieval effectiveness of collaborative tags and author keywords in different retrieval environments: An experimental study on medical collections. *Journal of the Association for Information Science and Technology*, 65(3), 483-500.

Mathes, A. (2004). *Folksonomies: Cooperative classification and communication through shared metadata*. Retrieved July 30, 2018 from http://www.bibsonomy.org/bibtex/245ae9616f7c7e480384d43cb2f6aec4d/jil.

Mohammad, F. (2018). Is preprocessing of text really worth your time for online comment classification? *ArXiv:1806.02908*. Retrieved July 30, 2018 from http://arxiv.org/abs/1806.02908.

Névéol, A., Doğan, R. I., & Lu, Z. (2010). Author keywords in biomedical journal articles. *AMIA Annual Symposium Proceedings*, 2010, 537-541.

Niwattanakul, S., Singthongchai, J., Naenudorn, E., & Wanapu, S. (2013). Using of Jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists, March 13-15, 2013*. Hong Kong.

Peters, I., Kipp, M. E. I., Heck, T., Gwizdka, J., Lu, K., Neal, D., & Spiteri, L. (2011). Social tagging & folksonomies: Indexing, retrieving... and beyond? *Proceedings of the 74th Annual Meeting of the American Society for Information Science and Technology*, 48(1), 1-4.

Rafferty, P. M. (2017). *ISKO Encyclopedia of knowledge organization: Tagging*. Retrieved July 30, 2018 from http://www.isko.org/cyclo/tagging.

Risueño, T. (2018). *What is the difference between stemming and lemmatization?* Retrieved August 29, 2018 from https://blog.bitext.com/what-is-the-difference-between-stemming-and-lemmatization/.

Stan, J., & Maret, P. (2017). Social bookmarking or tagging. In R. Alhajj, & J. Rokne (Eds.), *Encyclopedia of social network analysis and mining*. New York: Springer. https://doi.org/10.1007/978-1-4614-7163-9_91-1.

Strader, C. R. (2009). Author-assigned keywords versus Library of Congress Subject Headings: Implications for the cataloging of electronic theses and dissertations. *Library Resources & Technical Services*, 53(4), 243-251.

Syn, S. Y., & Spring, M. B. (2010). Tags as keywords: Comparison of the relative quality of tags and keywords. *Proceedings of the American Society for Information Science and Technology*, 46(1), 1-19.

Syn, S. Y., & Spring, M. B. (2013). Finding subject terms for classificatory metadata from user-generated social tags. *Journal of the American Society for Information Science and Technology*, 64(5), 964-980.

Thada, V., & Jaglan, V. (2013). Comparison of Jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology*, 2(4), 202-205.

Thomas, M., Caudle, D. M, & Schmitz, C. (2010). Trashy tags: Problematic tags in LibraryThing. *New Library World*, 111(5-6), 223-235.

United Nations Educational, Scientific and Cultural Organization. (2017). *The current status of ocean science around the world*. Paris: United Nations Educational, Scientific and Cultural Organization.

Voorbij, H. (1998). Title keywords and subject descriptors: A comparison of subject search entries of books in the humanities and social sciences. *Journal of Documentation*, 54(4), 466-476.

Wal, T. V. (2004). *You down with folksonomy?* Retrieved July 30, 2018 from http://www.vanderwal.net/random/entrysel.php?blog=1529.

Wu, D., He, D., Qiu, J., Lin, R., & Liu, Y. (2013). Comparing social tags with subject headings on annotating books: A study comparing the information science domain in English and Chinese. *Journal of Information Science*, 39(2), 169-187.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 39-51, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.4

Estimating the Impacts of Investment in a National Open Repository on Funded Research Output in South Korea

Hyekyoung Hwang
Korea Institute of Science and Technology Information, Seoul, Korea
E-mail: hkhwang@kisti.re.kr

Tae-Sul Seo
Korea Institute of Science and Technology Information, Seoul, Korea
E-mail: tsseo@kisti.re.kr

Yong-Hee Han
Department of Entrepreneurship and Small Business, Soongsil University, Seoul, Korea
E-mail: amade@ssu.ac.kr

Sung-Seok Ko*
Department of Industrial Engineering, Konkuk University, Seoul, Korea
E-mail: ssko@konkuk.ac.kr

ABSTRACT
Open access is a paradigm whereby the electronic versions of scholarly publications are made freely accessible without any restrictions. It is actively promoted globally and is also promoted domestically in accordance with this global trend. However, there is a growing need to evaluate existing activities and to seek policies for the steady spread of open access. This study examines the necessity of switching to a national repository from existing institutional repositories through policy direction analysis of open repositories. We examined domestic open access policies by analysing various overseas cases and the situation in South Korea. Finally, we determined the validity of investment in a national repository by analysing its social and economic impacts using the modified Solow-Swan model. The main parameters for applying the modified Solow-Swan model were estimated, and the domestic research and development expenditure was predicted via a regression method. Then, we applied a range of rate of returns to research and development (10% to 50%) to various scenarios and examined the effects of increasing accessibility and efficiency by 1% to 10%. We found that the implementation of a national open access repository in South Korea would have a substantial impact (to the tune of 147 billion won), without considering the potential costs of such a repository. Based on the estimates of the social and economic impact of a national repository, the implementation of a national open access repository in South Korea is economically viable. Besides having beneficial social and economic impacts, a national repository is expected to enhance awareness of open access among Korean researchers and institutions.

Keywords: open repository, open access, modified Solow-Swan model, national repository, economic analysis

Open Access

Accepted date: February 01, 2019
Received date: December 07, 2018

***Corresponding Author:** Sung-Seok Ko
Professor
Department of Industrial Engineering, Konkuk University, Seoul, Korea
E-mail: ssko@konkuk.ac.kr

1. INTRODUCTION

The scholarly communication ecosystem follows a clear cycle (Cox & Tam, 2018): It begins with research, followed by the production of an academic paper through organizing the research results, which is then distributed to others and used in other research (thereby producing new outcomes) following a process of evaluation, review, and publication in an academic journal. Traditionally, researchers have published their academic papers in academic journals created by an academic society or publishing company to disseminate their findings as well as to communicate and exchange opinions with other researchers. Many such academic journals have played a considerable role in the development of studies through the publication and distribution of academic papers. However, these types of journals have been criticized for a number of problems, such as the complex examination process, opacity of the publication process, high academic journal subscription fees, and publishers' abuse of copyright (Choi & Cho, 2005). Open access is one effort to resolve this situation. Open access is a new paradigm of academic information distribution, whereby anyone in the world can freely view academic research outcomes on the Internet. It represents an attempt to restore the essential characteristics of academic papers—opinion presentation, discussion, and idea sharing.

The full-scale implementation of open access is widely regarded as beginning with the 2002 Budapest Open Access Initiative (Budapest Open Access Initiative, 2002). Following this declaration, a push for open access began in earnest through the Bethesda Statement (2003) and Berlin Declaration (2003). Open access can be classified as gold open access and green open access, depending on the strategy (Schimmer, Geschuhn, & Vogler, 2015). Gold open access refers to the strategy of publishing an academic paper with open access in a journal that uses a peer review system. Gold open access papers are freely accessible at the time of publication online by anyone in the world. For an academic paper to be published as gold open access, it must be submitted to an open access journal (wherein all papers are published with open access) or a hybrid journal (wherein papers can be published under the subscription-based system or as open access, depending on the authors' choice). For a paper to be published as open access, either the authors themselves or a supportive organization must first pay an article processing charge (Lawson, 2016).

Green open access, on the other hand, is a strategy whereby authors self-archive their works in an open access repository or post them on their homepages, allowing anyone to freely access them online. There are various types of open access repositories, including institutional repositories (operated by the author's institution), subject repositories (operated by organizations for specific subject areas), and national repositories (operated by governmental bodies). When a paper is published using the green open access system, the paper becomes accessible in the repository only after a specific embargo period, which is set according to the license policy of the journal or publishing company. Green open access is a compromise between authors (who wish to publish their research papers in such a way that the papers will be widely read and quoted), users (who desire for papers to be freely available), and publishers (who want to profit from papers' sales) (Hwang, 2017).

Since the 2002 Budapest Open Access Initiative, numerous developed nations and prominent institutions have been actively pursuing open access policies. For example, the United States, the United Kingdom, and Spain have pushed for legal measures promoting the self-archiving style (i.e., green open access), while India, Denmark, Australia, and New Zealand have made efforts to promote open access via recommendations and encouragement systems. Major institutions in many of these countries have also shown tangible results in promoting open access through signing declarations and enacting policies (Kim, Kim, Choi, & Hwang, 2016). In the early 2010s, nearly a decade after the initial Budapest Open Access Initiative in 2002, there were several major reports conducting objective evaluations of and making suggestions on ensuring the sustainability of open access, which are widely regarded as a turning point for open access.

For instance, the Finch Report, published in 2012 by a working group composed of various interested parties (Finch, 2012), not only led to the development of the current UK open access policy base, but also had an influence on global open access policy, particularly that in Europe. This report examined ways of accelerating sustainable open access transition through cooperation with various interested parties involved in publishing and distributing research results, such as funding providers (research support institutions), universities, researchers, libraries, and publishers, while maintaining the basis of the scholarly communication ecosystem. The Finch report recommended a strategy of mixing gold and green open access to achieve a sustainable and orderly open access transition. Since the Finch report, there have been a number of important policy developments in the EU, the United States, and other major countries, and there have been several initiatives to actively support open access through forums such as Science Europe, the Global Research Council, and the G8. Particularly in the UK, there has been much greater progress in open access transition compared to other countries: About

19% of British publications are now published with gold open access, which is supported by various institutions that offer research funds (especially Wellcome Trust, Jisc, and the FP7 Pilot 'OpenAire' of the European Union) as well as a number of individual institutions. Most British universities have developed an institutional repository for green open access, which has resulted in rapid growth in repositories and the number of papers deposited therein (Tickell, 2016).

In 2015, a study by the Max Planck Digital Library (Schimmer et al., 2015) evaluating open access activities for over a decade argued that existing paid subscription journals should be converted into open access journals in order for open access to further proceed. For this to be achieved, the fund flow must be extensively restructured by converting from the existing subscription-based model to the gold open access model based on the article processing charge. This paper provided the theoretical base for the implementation of the OA2020. Led by the Max Planck Society in Germany, the OA2020 sought to covert at least 90% of existing subscription-based journals to gold open access journals by 2020. The OA2020 can be considered a cornerstone for the implementation of gold open access.

In South Korea, since the 2000s, universities and academic societies, particularly the Korea Institute of Science and Technology Information (KISTI) and National Library of Korea, have been taking steps to invigorate open access by pushing it in a limited number of fields and institutions. In addition, Open Access Korea (OAK) was formed in the early 2000s for managing public research results supported by the country's research and development (R&D) fund. OAK was composed of the OAK repository, Korea Journal Copyright Information, and OAK Central, which provides repository setup service to Korean academic journals. However, such open access papers written by Korean authors collected by OAK are mainly published in international journals. Therefore, open access remains largely at the level of collecting metadata and connecting these metadata to the original text because of the publisher's copyright on these papers (Hwang, 2017). Many researchers have emphasized the need for a national open access repository. For example, Seo, Heo, and Noh (2009) reported that there is a need for open access policies for public research results, starting with building field-specific open access repositories and providing greater cost support to manage a repository. In March 2009, the OAK project was implemented to promote open access and the common use and dissemination of knowledge. Furthermore, various policies, including the building of institutional repositories, have been pushed. Nevertheless, these efforts have largely failed in their intended purpose.

Given this situation—particularly that domestic public research results, especially in the fields of science and technology, are published mainly in overseas academic journals—it is necessary for South Korea to actively participate in international open access activities such as OA2020 and Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3), as well as to establish policies that focus on establishing open access repositories, in order to invigorate open access in Korea. SCOAP3 is an international collaboration in the high-energy physics community to convert traditional closed access physics journals to open access (SCOAP3 Consortium, 2019). Therefore, we examined the necessity of building a national open access repository via situational analysis and case studies to help invigorate adoption of green open access in South Korea. Furthermore, we executed a quantitative analysis on the potential economic and social effects of such a repository.

This paper is structured as follows: In the second section, we review the existing studies on open access repositories. In the third section, we describe the international and domestic situations of open access repositories, as well as the necessity of developing a national repository. We describe the quantitative analysis of the economic impact of implementing a national repository using Houghton's model in the following section. Finally, we describe the conclusions in the last section.

2. LITERATURE REVIEW

Previous research has explored the formation of open access repositories. Most of these studies examined how these repositories can support green open access, particularly in terms of user attitudes and behaviours (Kim, 2010), different disciplinary positions (Xia, 2007), and role changes for librarians (Walters, 2007).

There are also numerous practice-based case studies. Armbruster (2010) conducted a study on twelve repositories implemented in response to institutional or funder open access policies, while Davis and Connolly (2007) studied the reasons that end users accessed the Cornell University repository through faculty interviews and usage log files. Covey (2009) explained the attributes and behaviour of faculty who used the institutional repository at Carnegie-Mellon University. Koskinen et al. (2010) investigated the accommodation and usage of the institutional repository at the University of Helsinki. Roy, Biswas, & Mukhopadhyay (2012, 2013, 2016) investigated repositories in India.

To analyse the main characteristics of open access repositories and their global trends, most studies have employed

OpenDOAR data.[1] According to Morrison (2012), the number of repositories registered in OpenDOAR increased from 800 in 2006 to over 2,200 in 2012. Pinfield et al. (2014) analysed OpenDOAR data, focusing on global trends in open access repositories from 2005 to 2012. Wani, Gul, & Rah (2009) also analyzed OpenDOAR data between October 7 and 8, 2008, focusing on repository distribution by continent, country, core content type, operational status, software usage, repository type, subjects, and language. Abrizah, Noorhidawati, and Kiran (2017) analyzed state of repositories of Asian universities using OpenDOAR.

Green open access is considered relatively more accessible and cost-efficient than is gold open access. However, it is not as cheap as open access advocates initially claimed. Many education and research institutes, including universities, build and operate repositories using open source solutions such as DSpace and EPrints, but considerable construction and operational costs are incurred to ensure smooth utilization. Furthermore, there are other costs, such as verification costs for the copyrights of uploaded materials, costs related to correction of references, education costs for researchers, and operation costs, which differ according to the scale and the degree of utilization of the repository. For instance, Houghton et al. (2009) estimated that, assuming an author's uploading time is about 10 minutes, the cost for uploading papers to repositories in the UK is about 33 US dollars per paper. In the European Community-funded Publishing and the Ecology of European Research (2011), the cost of building an IT system for a full repository would be about 60,000 US dollars, while the personal cost per paper would vary substantially (2 to 53 US dollars, depending on the repository).

There have been various studies on the economic impacts of open access. The Research Information Network (2008) predicted that out of the total cost of journal publishing (25 billion pound), publishing costs and library costs account for 4.9 billion pound. Open access is estimated to be able to save 560 million pound. Houghton et al. (2009) estimated that open access would reduce the system cost of open access by about 212 million pound in the UK alone, with the greatest savings being for research performance (about 106 million pound). Houghton (2009), besides finding that open access would reduce system costs, found that the economic and social returns of open access to the UK's public-sector R&D would be about 170 million pound (based on the results of a modified version of the Solow-Swan model). He applied his model to Denmark and the Netherlands as well, and conducted a comparative analysis between these countries. He also estimated the economic and social impacts of the Federal Research Public Access Act in the US (Houghton, Rasmussen, & Sheehan, 2010).

3. INTERNATIONAL AND DOMESTIC OPEN ACCESS REPOSITORIES

3.1. Global Open Access Repositories

Since the development of DSpace and E-Prints in 2002, two major pieces of repository software, the construction of repositories has progressed in earnest. DSpace was jointly

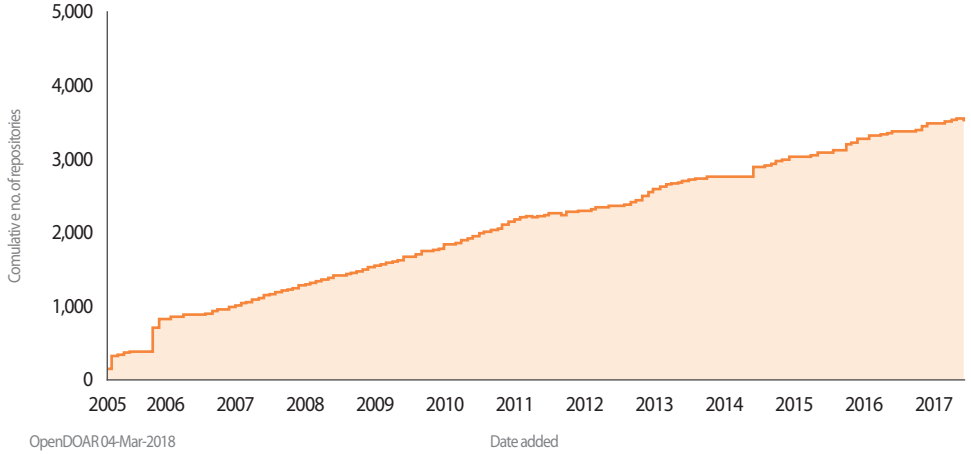


Fig. 1. Overall growth of repositories in OpenDOAR from December 2005 to March 2018.

[1] http://v2.sherpa.ac.uk/opendoar/

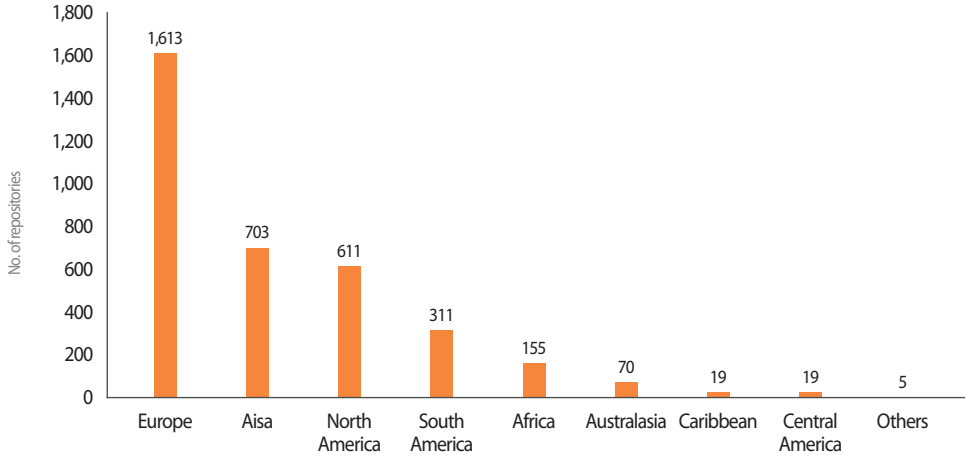


Fig. 2. Repository statistics by regions as of March 2018.

developed by the Massachusetts Institute of Technology and the HP Research Institute in the US and it quickly became known to the public because Cornell University utilized it to create its own repository. E-prints, developed by Southampton University in the UK, contributed substantially to the establishment and stabilization of the repository at Oxford University. Subsequently, the number of universities and research institutes worldwide that are developing repositories has steadily increased.

Fig. 1 shows the repository growth worldwide. The total number of repositories in OpenDOAR showed a steady increase (except in the first year) from 128 in December 2005 to 3,502 in March 2018. While there are slight differences in magnitude among regions, this increment was consistent across them. The increase can be attributed to growing awareness of open access.

Fig. 2 shows the repository statistics by regions as of March 2018. Europe had the highest number of repositories, at 1,162 (46% of the total), followed by Asia (702, 20%), North America (615, 18%), and South America (309, 9%). Asia—centred on Japan, India, Turkey, Indonesia, Taiwan, and China—is showing rapid growth in the number of repositories, to the point where the number recently surpassed that for North America. Therefore Asia, along with Europe, is becoming a centre of global open access repositories.

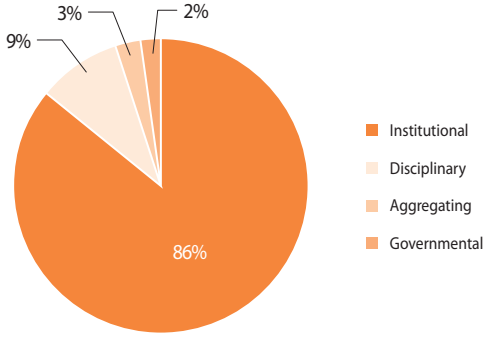


Fig. 3. Repository statistics by types as of March 2018.

When examining repository type (Fig. 3), most repositories were classified as institutional repositories (accounting for 86% of the total), followed by disciplinary repositories (at only 9% of the total). The proportion of institutional repositories is slowly increasing, indicating that recognition of open access is spreading and the number of requests for establishing institutional repositories is growing.

Table 1 compares the repositories of representative institutions and countries that are obliged to deposit public research results. The National Institutes of Health (NIH) in the United States (Organisation for Economic Co-Operation and Development, 2015) is in charge of depositing and utilizing research papers produced by the NIH fund in line with the national public deposit policy, while the Spanish Foundation for Science and Technology (Fundación Española para la Ciencia y la Tecnología, FECYT) operates a national repository called the 'Recolector de Ciencia Abierta' (RECOLECTA), based on a connection with the Network of Spanish University Libraries (Red de Biblioteca Universitarias Españolas). The Chinese Academy of Sciences (CAS, 2014) operates a repository called the 'CAS Institutional Repositories Grid' (CAS IR Grid) that comprehensively deposits and manages papers produced with CAS funds, linking them to the repositories of CAS-affiliated institutions.

Table 1. Comparison of repositories of representative institutions and countries that are obliged to deposit public research results

	PubMed Central	RECOLECTA	CAS IR Grid
Operator	NIH	FECYT, REBIUN	CAS
Object deposited	Research outcomes supported by NIH funds	Research outcomes supported by public funds	Research outcomes supported by CAS funds
Registration and depositor	Author or publisher	Author	Author
Connection system	NIH manuscript submission system	Institutional repositories	CAS-affiliated open access repositories
Form of materials	Final version accepted for journal publication	Edited version, preprint version	Final version of the paper, modified by author after peer review
Main service	Search, browsing	Search, browsing	Search, browsing
Main features	The provision of R&D statistical analysis data	Operates the National Open Access Repositories Community	Connection and integrated management at the original text level

RECOLECTA, Recolector de Ciencia Abierta; CAS IR Grid, Chinese Academy of Sciences Institutional Repositories Grid; NIH, National Institutes of Health; FECYT, Fundación Española para la Ciencia y la Tecnología [Spanish Foundation for Science and Technology]; REBIUN, Red de Biblioteca Universitarias Españolas [Network of Spanish University Libraries]; CAS, Chinese Academy of Sciences; R&D, research and development.

In February 2000, PubMed Central (PMC) built an open repository managed by the National Library of Medicine (NLM) which collects and stores papers published in biomedical and life science journals according to the NLM's legislative mandate for collecting and keeping biomedical papers. The academic journals fully participating in the PMC submit their papers to the PMC directly, and papers that are supported by NIH funds are directly deposited by the paper's author(s). In addition to its role as a repository, the PMC makes it possible to store and cross-reference data from various sources using a common format. Using the PMC, it is possible to find all related materials by quickly searching the entire collection of full-text documents. The PMC also integrates literatures from different fields in order to improve the research and knowledge of experts such as scientists and clinicians. As of March 2018, about 4.7 million articles from approximately 7,000 journals are retained in the PMC, and the number of fully participating journals is 2,098.

In 2007, the Spanish government encouraged the establishment of an open access repository, announcing the 'Draft of the National Law of Science.' This law included a regulation whereby researchers who received public funds had to make their research results open access within six months. Article 37 of Spanish Law 14/2011 on Science, Technology, and Innovation (named 'Open Access Dissemination') established a national standard stipulating that the outcomes of research activities supported by the state must be deposited in open repositories. Furthermore, since 2007, Spain's FECYT and Red de Biblioteca Universitarias Españolas have sought to build a national infrastructure for open repositories; accordingly, through steady collaboration, these two organizations conceived RECOLECTA, an open platform that links all institutional open repositories in Spain and provides services for repository managers, researchers, and decision makers. RECOLECTA has promoted and coordinated a national infrastructure for interoperable digital science repositories utilizing standards adopted by communities worldwide and was designed to promote research development and the adoption of open policy. Specifically, the RECOLECTA provides easy, free access to all scientific research outcomes stored in Spain's repositories, as well as seeking to build, maintain, support, and improve the national repository infrastructure. Specifically, it provides users with support services, enhances the national open community, and offers statistical data on repositories.

The CAS is a core pioneering organization in the field of Chinese technology and natural sciences, consisting of a comprehensive R&D network, higher education system, and outcome-based academic society. In China, open access began in 2003 through participation of Chinese scholars in open access and academic publishing seminars. Open access only became standardized the following year when the CAS and National Natural Science Foundation of China signed the Berlin Declaration. Since then, China has been constantly working on open-access-related activities, such as establishing a CAS institutional repository system (CAS IR Grid) on a trial basis in 2007 and opening access to China's information portal in 2008. The CAS IR Grid contains 114 institutional repositories as of March 2018. When a researcher deposits his or her paper into a CAS-affiliated institutional repository, the paper becomes available in the CAS IR Grid. If an institution does not yet have a repository, papers must be deposited in the repository operated by the National Science Library of the CAS. Since 2012, the annual number of papers registered in the CAS IR Grid has ranged from 40,000 to 1,500,000, of which more than 70% have the original text available. As of March 2018, there are about

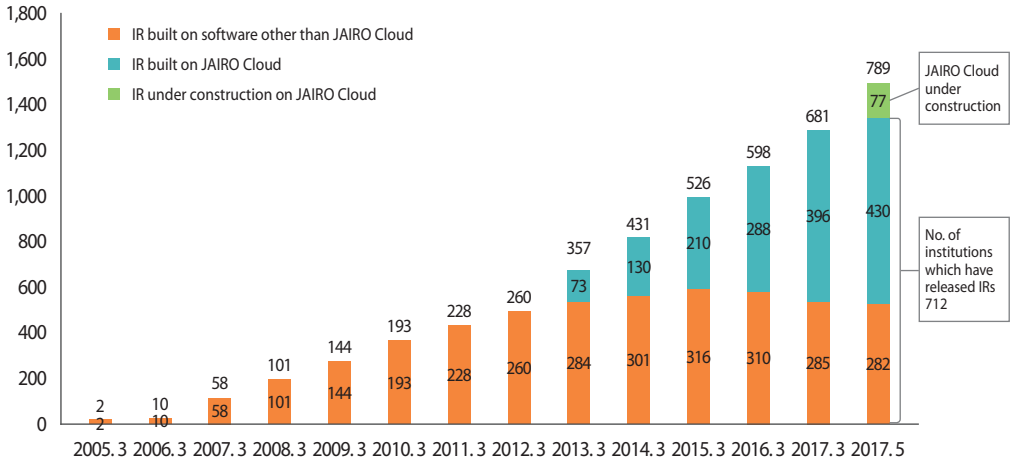


Fig. 4. Number of institutional repositories in Japan as of August 2017. IR, institutional repository.

820,000 registered papers, of which about 75% provide the original text.

Japan began its Cyber Science Infrastructure Program in 2005. Under this program, institutions (including universities) have begun devoting some effort to establishing their own repositories. In fact, by 2010 nearly 200 institutions across Japan had established an institutional repository. Research institutions, including universities, independently established these repositories using software such as DSpace. However, many institutions, while expressing a willingness to build a repository, found it difficult to afford or hesitated because of the expected burden of operating the repository after establishing it.

Based on past experiences of establishing a repository, the National Institute of Informatics of Japan promoted the introduction of the JAIRO Cloud in 2011 in order to promote wider development of repositories. The JAIRO Cloud, established in 2012, is a computing service based on an SaaS system that was created by the National Institute of Informatics to improve the operation of institutional repositories. Initially it targeted universities without an institutional repository, but since January 2014 it has begun accepting institutions with an existing repository. In May of that same year, starting with the transition of Tulips-R (the institutional repository of the University of Tsukuba), established institutional repositories throughout Japan began transferring to JAIRO Cloud. Fig. 4 shows the state of establishment of institutional repositories in Japan, and it is evident that rapid growth has been achieved as a result of introduction of the JAIRO Cloud, with many institutions transferring from their own repository to the JAIRO Cloud.

3.2. Korean Open Repositories

The establishment of institutional repositories in South Korea began with the Korea Advanced Institute of Science and Technology (KAIST) Open Access Self-Archiving System (KOASAS). In 2007, KAIST allocated some of its own budget to develop and operate a repository for managing, preserving, and distributing research outcomes obtained by the professors and researchers of the university. KOASAS utilizes the same model operated by the libraries. Later, in February 2012, KAIST made active use of its institutional repository by establishing the Researcher Information Management System, a performance evaluation system of researchers in KAIST, and connected it with KOASAS. As KOASAS holds more than 200,000 academic articles, including papers published in 2018, it is a valuable resource for researchers in South Korea and abroad.

The central library of Seoul National University officially launched its institutional repository S-Space in December 2008.[2] This repository was developed by benchmarking with DSpace and KOASAS, upgrading and customizing for the convenience of its members. More than 98,000 materials have been registered in S-Space as of March 2018, including research papers published in academic journals, papers presented at academic conferences, and dissertations issued by various academic societies and institutions affiliated with Seoul National University. In 2017 only, there were over 6 million downloads.

The full-scale implementation of a domestic open access

[2] http://s-space.snu.ac.kr/

repository was the OAK Project. The OAK Project, which was promoted by KISTI in March 2009, sought to build knowledge cooperation to promote open access for domestic academic information by adhering to the following steps: repository development and dissemination, open access journal publication support, open access portal (OAK Central) establishment, and open access governance system establishment. Following the replacement of the host organization by the National Library of Korea in 2014, the development and dissemination of the Korean OAK repository began.

The OAK repository was built using DSpace, and customized to the domestic environment. It was distributed through the help of an OAK repository operation consultative group, consisting of KISTI, repository system developers (KISTI's partners), and the OAK repository operation organization. This consultative group selects target institutions for new repository establishment through a public contest and shares the trends and know-how in operation through training and seminars with the selected institutions. About five institutions are selected annually based on their applications for establishing the repository; as of March 2018, a total of 38 OAK repositories have been established. KISTI's partners played a role in spreading OAK repositories to various institutions that wished to install it. Furthermore, the OAK repository is continually updated, in accordance with updates to DSpace.

In addition, institutions that installed the OAK repository identified new requirements through operation of the repository, thus helping the consultative group to improve the repository and its operation methods. The contents of all OAK repositories can be retrieved through integrated search services by both domestic and foreign users through the OAK portal, which is operated by the National Library of Korea. Interested users can access the original texts of content via the repository portal. This has helped increase web traffic to the OAK repository.

However, not all institutional repositories in South Korea are smoothly operated. While there are about 24 unique repositories that hold academic papers, only a few—such as KOASAS and S-Space—are actively operating. Considering the number of domestic universities and public institutions, this figure indicates exceedingly poor performance compared with Europe and Japan.

Open access is being pushed in various directions all over the world. For example, gold open access is being implemented through such policies as the OA2020 (led by the Max Planck Digital Library), SCOAP3 (centred on European Council for Nuclear Research), and the Big Deal models of various European countries, whereas green open access is being implemented through the establishment of open access repositories. Although South Korea is making a considerable effort to keep up with this trend, its achievements are comparatively limited because of problems such as peculiar characteristics in the domestic academic ecosystem, limited participating institutions, lack of government policy support, and low awareness among researchers of open access (Hwang, 2017).

As part of an effort to overcome this problem, the OAK Project is seeking to promote collaboration among institutions in the development of repositories through OAK Central. However, there is still a need to build a national repository such as Japan's JAIRO Cloud and Spain's RECOLECTA. The current situation in South Korea is similar to that in Japan before the introduction of JAIRO Cloud. In particular, while a number of institutions have established an institutional repository, some are in name only, as the institutions are incapable of maintaining their operation. It is therefore necessary to implement a nationally integrated repository, as well as to build up personalized institutional repositories for institutions which desire to build repositories, but lack the capability as well as the necessary technology to do so.

4. ESTIMATED IMPACT OF A NATIONAL REPOSITORY IN SOUTH KOREA

4.1. Model Outline

It is difficult to calculate the potential impact of implementing an open access repository, and doing so can cause a considerable degree of controversy. Nevertheless, to assess potential impacts and use them for future reference, Houghton developed a model using the Solow-Swan model (for further detail, refer to Houghton et al., 2009).

The basic Solow-Swan model (Solow, 1957) is represented in the following production function:

$$Y = A^{\eta} K^{\beta} L^{\alpha}$$

where A is an index of technology, K is the capital stock, and L is the supply of labour. Both K and L are taken to be fully employed by virtue of the competitive markets assumption. Solow further developed this model, proposing that once we exclude the impacts of capital and labour, what is left is the impact of technology. He subsequently studied the impact of technological development on overall production. He also applied the model to estimate the rate of return to R&D.

This model is based on several major assumptions. The first assumption is that all R&D creates useful knowledge in

economic or social terms (the efficiency of R&D). The second assumption is that all created knowledge is equally accessible to anyone who wants to use it for productive activities (accessibility of knowledge).

However, in the real world there are numerous barriers or limitations to accessing and utilizing knowledge. Based on this, Houghton (2009) demonstrated that it is possible to calculate the impact on return to R&D by improving the accessibility and efficiency of knowledge and reducing friction. In this modified Solow-Swan model, accessibility and efficiency are considered 'friction variables.' He proposed the following formula:

$$\frac{\partial y}{\partial R} = \Upsilon \frac{Y}{R}(1+\delta_{\phi})(1+\delta_{s})$$

Where δ_{ϕ} $(1+\delta_{s})$ is the percentage change in efficiency (accessibility), Υ represents the contribution ratio of the rate of growth of R&D knowledge stock to output growth as a factor of production (i.e., the elasticity), and R indicates the stock of R&D knowledge, which can be calculated as follows:

$$R_t = (1-\delta)R_{t-1} + R\&D_{t-1}$$

where δ is the rate of obsolescence of the knowledge stock.

4.2. Operationalizing the Model

The main parameters for applying the modified Solow-Swan model are rate of return to R&D, accessibility, and efficiency. Research on the economic impact of R&D at the firm, industry, and national levels has been increasing. However, the claimed variation in the rate of return to R&D differs widely among researchers. For example, Salter and Martin (2001) found that the rates ranged from 10% to 150%. Hall, Mairesse, and Mohnen (2010) found similar degrees of variation depending on the researcher and analysis level. When all results of these studies are combined, a conservative estimate puts the rate of return to R&D at between 10% and 20%.

Accessibility can be defined as the proportion of the stock of knowledge generated by R&D accessible to those who would use it productively. Houghton et al. (2010) suggested measuring the increment in accessibility by combining the degree of access to desired academic information (access gaps), the degree at which academic information was cited (citation), and the degree of variation at which academic information was downloaded. Although the degree differs according to the characteristics of the repository to be built, the results of existing studies suggest that accessibility can be increased by as much as 4.5% (as a conservative estimate).

Efficiency can be defined as the proportion of R&D spending that generates useful knowledge; it can have a number of dimensions relating to wasteful, inefficient, and/or poorly directed research expenditures. Houghton et al. (2010) suggested using scenario-based measurement tools for efficiency, such as wasteful expenditure, number of new opportunities, and time saving for research.

Various other parameters must also be defined. First, a project to establish a research repository can be considered a kind of 'information system business.' Considering the life-cycle of a system in South Korea, the analysis period of the main information system is generally four to seven years. Therefore, we conducted a study spanning five years, which falls in the middle of this range.

There is a time lag between research spending and the social and economic impact of research results. In some fields, this lag can range from 2 to 30 years or more, whereas in others, the lag is no more than 1 to 2 years. According to Mansfield (1991, 1998), the average lag in US firms between publication of academic research and the timing of a related commercial innovation was around 7 years (which fell to 6.2 in the later study). Adding the time for publication, the lag was about 10 years, but it can be assumed that the time was shortened when considering the difference from the time when the research results were announced. Accordingly, this study assumed a lag time of 7 years.

Since the cost and benefits of a business manifest over a long period of time, it is necessary to compare them by converting all the costs and benefits that will occur in the future to their present value. This conversion process means discounting the current value, and the interest rate applied at this time is called

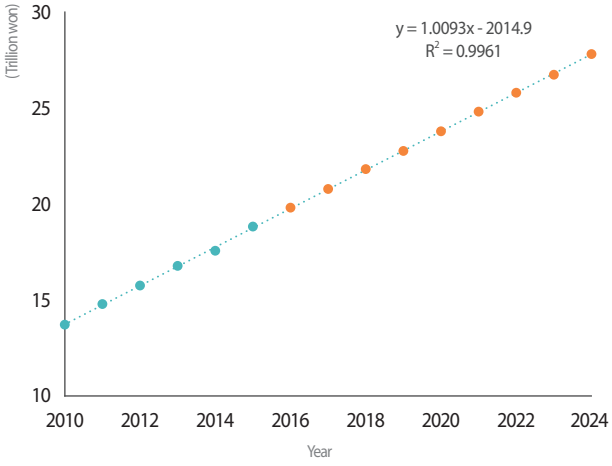


Fig. 5. Scale and prediction of research and development expenditure.

the discount rate. Since the task of estimating the appropriate social discount rate is exceedingly complicated, domestic studies in general use the 5.5% social discount rate presented by the Korea Development Institute.

Fig. 5 shows the results of estimating domestic R&D expenditure from 2016 to 2024 via a regression analysis, drawing on information of domestic R&D expenditure from 2010 to 2015. While the structure of the R&D expenditure is decided by policy, its coefficient of determination (adjusted R^2) was nevertheless very high; thus, it can be considered a valuable estimate. The results indicated the estimated R&D expenditure in 2019 and 2024 would be 22.9 trillion won and 28 trillion won, respectively, with an estimated annual growth rate of 4.4%.

Table 2 shows the estimates of the impacts of a national open access repository. For illustrative purposes, we expanded the range of rate of returns on R&D from 10% to 50% so that it could be applied to various scenarios. We also examined the increases in accessibility and efficiency by 1% to 10%.

With a 20% return to R&D expenditure on 22.9 trillion won in 2019, an increase of about 1% in accessibility and efficiency yields a return to R&D of about 63 billion won. This is a discounted amount based on 2019, taking into account the 7 years of time-lag between expenditure and impact. Overall, it is evident that the increase in R&D expenditure leads to an increase in impact. The increasing rate of return on R&D is beyond the rate of increase in R&D expenditure. The increase in accessibility and efficiency also appears to have a strong influence on the impact.

Table 2. Estimates of the impact of investment in a national open repository

		Rate of return on R&D (billion won)				
		10%	20%	30%	40%	50%
2019 (22,873 billion won)						
Percent change in accessibility & efficiency	1%	32	63	95	127	158
	2%	64	127	191	255	318
	5%	161	323	484	646	807
	10%	331	662	993	1,323	1,654
2020 (23,882 billion won)						
Percent change in accessibility & efficiency	1%	33	66	99	132	165
	2%	66	133	199	266	332
	5%	169	337	506	674	843
	10%	345	691	1,036	1,382	1,727
2021 (24,891 billion won)						
Percent change in accessibility & efficiency	1%	34	69	103	138	172
	2%	69	139	208	277	346
	5%	176	351	527	703	879
	10%	360	720	1,080	1,440	1,800
2022 (25,901 billion won)						
Percent change in accessibility & efficiency	1%	36	72	108	143	179
	2%	72	144	216	288	360
	5%	183	366	549	731	914
	10%	375	749	1,124	1,499	1,873
2023 (26,910 billion won)						
Percent change in accessibility & efficiency	1%	37	75	112	149	186
	2%	75	150	225	300	374
	5%	190	380	570	760	950
	10%	389	779	1,168	1,557	1,946

R&D, research and development.

Table 3. NPV of estimates of the impact of investment in a national open repository

NPV (base: 2018)		Rate of return on R&D (billion won)				
		10%	20%	30%	40%	50%
Percent change in accessibility & efficiency	1%	147	293	440	586	733
	2%	295	589	884	1,178	1,473
	5%	747	1,494	2,242	2,989	3,736
	10%	1,531	3,062	4,593	6,123	7,654

NPV, net present value; R&D, research and development.

Table 2 shows the increasing impacts over the years, as well as the result of converting them into the present value (2018) for basic economic analysis. As mentioned above, this refers to the value calculated for every 5 years based on the economic lifecycle; it could be much more effective if the lifecycle were longer than 5 years. Therefore, the values presented in Table 2 show the social and economic impacts of the establishment of a national open repository. This provides a guideline for investment in the establishment of a national open access repository. When focusing on the most conservative situation, if the rate of return on R&D is 10% and the rate of increase in accessibility and efficiency is 1%, there is still an impact of about 147 billion won. By contrast, in the moderate situation (30% rate of return to R&D, 5% increase in accessibility and efficiency), the impact is over 2.2 trillion won.

In this study, we exclude the costs of establishing and operating the national open repository. This is because these would differ considerably according to the architecture and scope of application of the to-be-established system, and estimating without reliable information on system design is foolhardy at best. Nevertheless, the information in Table 3 provides a rough guideline for national open repository investment.

5. CONCLUSION

There is plenty of research on the necessity of open access, so much so that it is often taken for granted by researchers and policymakers. Open access is being actively promoted around the world. The OA2020, SCOAP3, and various Big Deal models in European countries have demonstrated a new direction for gold open access and are producing important results with support by numerous researchers and institutions. However, presently gold open access has a somewhat limited scope in terms of the types of academic papers published. Therefore, implementing gold open access in earnest on a global scale requires more time. As an alternative to this, green open access, which involves the use of open access repositories, has received steadily increasing attention. Research on the establishment of such repositories was initially centred on Europe and North America, but is now actively being conducted in Asian countries, mainly Japan, China, India, and Indonesia.

Looking at the Korean situation, the establishment of institutional repositories under the OAK Project is continuing. However, the performance of this project is falling short of expectations because of limitations in managing already established repositories. To overcome this issue, it is now necessary to promote an alternative to these institutional repositories by implementing a national repository, as in the case of Japan and Spain. Our calculation of the social and economic impact of such a repository by applying Houghton's modified Solow-Swan model revealed that a national repository would have an impact of 147 billion won, even when using a conservative approach. Although we did not perform a comprehensive analysis of the potential costs, these findings are nevertheless encouraging for South Korea. Besides the social and economic impacts, a national repository is expected to enhance awareness of open access among Korean researchers and institutions.

REFERENCES

Abrizah, A., Noorhidawati, A., & Kiran, K. (2017). Global visibility of Asian universities' open access institutional repositories. *Malaysian Journal of Library & Information Science*, 15(3), 53-73.

Armbruster, C. (2010). *Implementing open access: Policy case studies*. Retrieved December 7, 2018 from http://papers.ssrn.com/sol3/papers.cfm?abstract_id= 1685855.

Chinese Academy of Sciences (2014). Chinese Academy of Sciences policy statement on open access to articles from

publicly funded scientific research projects. *Bulletin of the Chinese Academy of Sciences*, 28(3), 230.

Budapest Open Access Initiative (2002). *Read the Budapest Open Access Initiative*. Retrieved December 7, 2018 from http://www.budapestopenaccessinitiative.org/read.

Choi, J. H., & Cho, H. Y. (2005). The recent trends of open access movements and the ways to help the cause by academic stakeholders. *Journal of the Korean Society for Information Management*, 22(3), 307-326.

Covey, D. T. (2009). Self-archiving journal articles: A case study of faculty practice and missed opportunity. *portal: Libraries and the Academy*, 9(2), 223-251.

Cox, A. M., & Tam, W. W. (2018). A critical analysis of lifecycle models of the research process and research data management. *Aslib Journal of Information Management*, 70(2), 142-157.

Davis, P. M., & Connolly, M. J. L. (2007). Institutional repositories: Evaluating the reasons for non-use of Cornell University's installation of DSpace. *D-Lib Magazine*, 13(3/4). doi:10.1045/march2007-davis.

Finch, J. (2012). *Accessibility, sustainability, excellence: How to expand access to research publications. Report of the Working Group on Expanding Access to Published Research Findings*. London: Home Office.

Hall, B. H., Mairesse, J., & Mohnen, P. (2010). Measuring the returns to R&D. In B. H. Hall, & N. Rosenberg (Eds.), *Handbook of the economics of innovation* (pp. 1033-1082). Amsterdam: North-Holland Publishing.

Houghton, J. (2009). *Open access: What are the economic benefits? A comparison of the United Kingdom, Netherlands and Denmark*. Retrieved December 7, 2018 from https://ssrn.com/abstract=1492578.

Houghton, J., Rasmussen, B., & Sheehan, P. (2010). *Economic and social returns on investment in open archiving publicly funded research outputs*. Washington, DC: Scholarly Publishing & Academic Resources Coalition. Retrieved December 7, 2018 from http://sparc.arl.org/sites/default/files/vufrpaa.pdf.

Houghton, J., Rasmussen, B., Sheehan, P., Oppenheim, C., Morris, A., Creaser, C., Gourlay, A. (2009). *Economic implications of alternative scholarly publishing models: Exploring the costs and benefits. A report to the Joint Information Systems Committee (JISC)*. Melbourne: Victoria University. Retrieved December 7, 2018 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.5031&rep=rep1&type=pdf.

Hwang, H G. (2017). Concept of open access and policy trend. *IE Magazine*, 24(4), 24-29.

Kim, J. (2010). Faculty self-archiving: Motivations and barriers. *Journal of the American Society for Information Science and Technology*, 61(9), 1909-1922.

Kim, S.-Y., Kim, J., Choi, H., & Hwang, H. (2016). An analysis on open access policies on publications funded by overseas public institutions. *Journal of the Korean Library and Information Science Society*, 50(4), 209-229.

Koskinen, K., Lappalainen, A., Liimatainen, T., Niskala, A., Salminen, P.J., & Nevalainen, E. (2010). The current state of open access to research articles from the University of Helsinki. *ScieCom Info*, 6(4), 1-7.

Lawson, S. (2016). Report on offset agreements: Evaluating current JISC Collections deals. Year 1--evaluating 2015 deals. Retrieved December 7, 2018 from https://dx.doi.org/10.6084/m9.figshare.4047777.v1.

Mansfield, E. (1998). Academic research and industrial innovation: An update of empirical findings. *Research Policy*, 26(7/8), 773-776.

Max Planck Institute (2003). *Berlin declaration on open access to knowledge in the sciences and humanities*. Retrieved December 7, 2018 from https://openaccess.mpg.de/Berlin-Declaration.

Morrison, H. (2012, October 6). Thank you, open access movement! September 30, 2012 Dramatic growth of open access [Web log post]. Retrieved December 7, 2018 from https://poeticeconomics.blogspot.com/2012/10/thank-you-open-access-movement.html.

Publishing and the Ecology of European Research. (2011). *PEER economics report*. Milano: Universit Bocconi. Retrieved December 7, 2018 from http://www.peerproject.eu/fileadmin/media/reports/PEER_Economics_Report.pdf.

Pinfield, S., Salter, J., Bath, P. A., Hubbard, B., Millington, P., Anders, J. H., & Hussain, A. (2014). Open-access repositories worldwide, 2005-2012: Past growth, current characteristics, and future possibilities. *Journal of the Association for Information Science and Technology*, 65(12), 2404-2421.

Research Information Network. (2008). *Activities, costs and funding flows in the scholarly communications system in the UK: Report commissioned by the Research Information Network*. Retrieved December 7, 2018 from http://www.rin.ac.uk/system/files/attachments/Activites-costs-flows-report.pdf.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2012). Study of open access repositories: a global perspective. In *Information-Innovation-Technology: Creating Seamless Linkages, 29th Convention & Conference of Society*

of Information Science. Silchar: National Institute of Technology.

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2013). Global visibility of Indian open access institutional digital repositories. *International Research: Journal of Library and Information Science*, 3(1).

Roy, B. K., Biswas, S. C., & Mukhopadhyay, P. (2016). Open access repositories for Indian universities: towards a multilingual framework. *IASLIC Bulletin*, 61(4), 150-161.

Salter, A. J., & Martin, B. R. (2001). The economic benefits of publicly funded basic research: A critical review. *Research Policy*, 30(3), 509-532.

Schimmer, R., Geschuhn, K. K., & Vogler, A. (2015). *Disrupting the subscription journals' business model for the necessary large-scale transformation to open access*. München: Max Planck Digital Library.

SCOAP3 Consortium. (2019). *What is SCOAP3?* Retrieved December 7, 2018 from https://scoap3.org/what-is-scoap3/.

Seo, T. S., Heo, S., & Noh, K. R. (2009). *Public access policy for scholarly journal open access: KISTI knowledge report 4*. Daejeon: Korea Institute Science and Technology Information.

Solow, R. M. (1957). Technical change and the aggregate production function. *Review of Economics and Statistics*, 39(3), 312-320.

Tickell, A. (2016). *Open access to research publications: Independent advice*. London: Department for Business, Innovation & Skills.

Wani, Z. A., Gul, S., & Rah, J. A. (2009). Open access repositories: A global perspective with an emphasis on Asia. *Chinese Librarianship: An International Electronic Journal*, 27. Retrieved December 7, 2018 from http://www.iclc.us/cliej/cl27WGR.pdf.

Walters, T. O. (2007). Reinventing the library: How repositories are causing librarians to rethink their professional roles. *portal: Libraries and the Academy*, 7(2), 213-225.

Xia, J. (2007). Assessment of self-archiving in institutional repositories: Across disciplines. *Journal of Academic Librarianship*, 33(6), 647-654.

JISTaP http://www.jistap.org
Journal of Information Science Theory and Practice
eISSN : 2287-4577 pISSN : 2287-9099

Research Paper
J Inf Sci Theory Pract 7(1): 52-71, 2019
https://doi.org/10.1633/JISTaP.2019.7.1.5

Unified Psycholinguistic Framework: An Unobtrusive Psychological Analysis Approach Towards Insider Threat Prevention and Detection

Sang-Sang Tan*
Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore
E-mail: tans0348@ntu.edu.sg

Jin-Cheon Na
Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore
E-mail: tjcna@ntu.edu.sg

Santhiya Duraisamy
School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
E-mail: santhiya003@ntu.edu.sg

ABSTRACT
An insider threat is a threat that comes from people within the organization being attacked. It can be described as a function of the motivation, opportunity, and capability of the insider. Compared to managing the dimensions of opportunity and capability, assessing one's motivation in committing malicious acts poses more challenges to organizations because it usually involves a more obtrusive process of psychological examination. The existing body of research in psycholinguistics suggests that automated text analysis of electronic communications can be an alternative for predicting and detecting insider threat through unobtrusive behavior monitoring. However, a major challenge in employing this approach is that it is difficult to minimize the risk of missing any potential threat while maintaining an acceptable false alarm rate. To deal with the trade-off between the risk of missed catches and the false alarm rate, we propose a unified psycholinguistic framework that consolidates multiple text analyzers to carry out sentiment analysis, emotion analysis, and topic modeling on electronic communications for unobtrusive psychological assessment. The user scenarios presented in this paper demonstrated how the trade-off issue can be attenuated with different text analyzers working collaboratively to provide more comprehensive summaries of users' psychological states.

Keywords: insider threat, psycholinguistics, text analysis, sentiment analysis, emotion analysis, topic modeling

Open Access

Accepted date: March 15, 2019
Received date: February 04, 2019

***Corresponding Author:** Sang-Sang Tan
PhD Candidate
Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, 637718, Singapore
E-mail: tans0348@ntu.edu.sg

1. INTRODUCTION

Compared to the advancements in outsider threat prevention and detection, the development of solutions for monitoring insider threat is still in its early stage. The discrepancy in the state of the art of outsider and insider threat mitigation does not necessarily mean that outsider threat poses greater risks to organizations. In fact, with their legitimate and privileged access to an organization's assets and their knowledge of the internal workings of the organization, it is easier for malicious insiders to target the vulnerabilities of the organization without having to overcome most of the barriers that protect the organization against outsiders. Therefore, adversarial insiders have the potential to cause more damage than outside attackers. To make matters worse, insiders are also in a better position to cover their tracks and to perpetrate crimes without being detected.

There is a growing body of literature that recognizes the importance of protecting organizations against insider threat. Gheyas and Abdallah's (2016) systematic literature review and meta-analysis revealed a clearly discernible upward trend in the number of publications related to insider threat mitigation from the year 2000. In general, many studies in this line of research including those of Chen and Malin (2011), Eberle, Graves, and Holder (2010), and Myers, Grimaila, and Mills (2009) have confined the scope of insider attacks to malicious activities that occur in the computational environment, such as data sabotage and espionage happening over organizational computing systems and networks. In the present study, however, we have taken a more general approach that views insider attacks as all types of malicious acts taken by anyone who has access to organizational resources, facilities, and information that would put an organization at risk or cause the organization to suffer any forms of loss. Examples of these malicious conducts include, but are not limited to, scenarios in which a trusted partner with legitimate access to organizational data secretly provides the data to the organization's competitor, or a former employee passes on sensitive information of an organization to his new employer. Our general view of insider attacks has an important implication: Given that we make no assumption of the context or environment in which these malicious acts might be carried out, commonly used methods that focus on tracking activities in organizational systems and networks may not be applicable. Specifically, the fact that some insider crimes might not involve unauthorized access or other anomalous conduct that can draw suspicion to an insider makes them difficult to detect through activity tracking. With respect to this difficulty, we posit that the mitigation of insider threat should also give emphasis to analyzing internal states like the personalities and emotions of individuals in addition to tracking external acts.

When it comes to modeling insider threat, there are numerous theoretical models to draw upon, the most commonly referred to being the generic set of Capability-Motivation-Opportunity (CMO) models (Schultz, 2002) that describes insider threat as a function of three dimensions: motivation, opportunity, and capability. As pointed out by Colwill (2009), while various technical and procedural solutions are available to address issues related to opportunity and capability, assessing motivation is usually more challenging. One's motivation to commit a malicious act is often affected by internal factors such as personalities and emotional states, which need to be assessed through psychological analysis. However, a direct psychological examination is not always an option considering the legal, ethical, and privacy concerns that might arise from this practice (Brown, Greitzer, & Watkins, 2013; Greitzer, Frincke, & Zabriskie, 2010; Kiser, Porter, & Vequist, 2010). Moreover, this kind of assessment is also obtrusive in nature and might be perceived as unfounded accusation and scrutiny, thus running the risk of causing human conflicts in organizations.

To be able to perform psychological analysis while minimizing the aforementioned risks, more recent attention has focused on automated text analysis of electronic communications as an alternative approach for carrying out behavior monitoring (Brown, Greitzer, & Watkins, 2013; Brown, Watkins, & Greitzer, 2013). These works originated from the field of psycholinguistics, an interdisciplinary field that studies the interrelation between psychological and linguistic aspects. Earlier research in psycholinguistics has shown that language use is correlated with psychological and emotional states (Pennebaker, Booth, & Francis, 2001; Pennebaker, Mehl, & Niederhoffer, 2003). As Brown, Greitzer, and Watkins (2013) noted, the primary advantage of using this kind of psycholinguistic approach is that "…organizations may unobtrusively monitor any and all individuals who routinely generate text with the organizations' information systems. As human analysts are excluded from the early phases of such analysis, the psycholinguistic approach may provide a means of monitoring psychosocial factors in a uniform and non-discriminatory manner" (p. 2).

Although existing studies on psycholinguistics have provided a good start, they have yet to reach the advancement needed for making a significant impact on predicting and detecting insider attacks. A major impediment to the progress of this line of research is the dilemma in balancing the risk of missed catches and the number of false alarms. On one hand, in order to minimize the risk of missing any potential threat, an optimal text analyzer should report all suspicious signs of insider threat.

On the other hand, the false alarm rate should not be too overwhelming for the output of the analysis to be actionable. This difficulty is complicated by the imperfection of text analysis methods which, in their current state of the art, cannot provide highly accurate results necessary for achieving these goals.

In the interest of finding the sweet spot in dealing with the trade-off between two seemingly contradictory states, we propose a unified psycholinguistic framework that combines multiple text analysis methods including sentiment analysis, emotion analysis, and topic modeling for unobtrusive psychological assessment. Both sentiment analysis and emotion analysis are fast-growing research areas in affective computing, a field focusing on the development of technology that enables machines to recognize and process human affect. The key difference between these two types of analyses is that the former refers to the recognition of sentiment valences (positive, neutral, or negative) whereas the latter embraces a more fine-grained analysis of human emotions (such as anger, joy, sadness, etc.). Finally, topic modeling complements our framework by facilitating the identification of significant topical patterns from the textual data.

The objective of this work is thus to develop and demonstrate the viability of a framework that combines methodologically diverse text analyzers to analyze insiders' written communications and monitor their psychological and emotional states. The proposed framework has a twofold bearing upon minimizing the risk of missed catches while maintaining a low false alarm rate. First, by taking into consideration the outputs generated by multiple text analyzers, the uncertainty in tracking potentially malicious insiders can be greatly reduced. For instance, when the outputs of all analyzers suggest that an employee has shown absolutely no sign of threat, security analysts can more confidently exclude the employee from the list of suspicious individuals. Likewise, if all analyzers indicate that an employee has a high potential of becoming a threat, it would seem reasonable to keep an extra eye on this employee or to take preventive actions that minimize the possibility of any future wrongdoings. Therefore, having multiple analyzers provides more assuring evidence to either dismiss or support further investigation. Second, each analyzer in the framework might be superior in some cases but less so in others. These text analyzers can thus complement each other in the sense that one analyzer might be able to capture the signs that other analyzers have missed. In particular, due to the fundamental technical limits embodied in different text analysis methods, different analyzers might generate contradictory results in the assessment of the same individual. With a unified framework in which different text analyzers work collaboratively across methodological divides, these contradictory results can serve as the indicator for invoking further investigation, thus reducing the risk of overlooking any signs of potential threat.

2. RELATED WORK

2.1. Conceptual Modeling of Insider Threat

Schultz (2002) pointed out that many conceptual models of insider threat can be subsumed under the broader umbrella of CMO models. Variants of CMO models include those described by Parker (1998) and Wood (2000). In general, the CMO models suggest that insider attacks happen with the presence of the following essential components (Schultz, 2002):

- Capability, which refers to the level of relevant knowledge and skill that would enable an insider to commit the crime
- Motivation, which encompasses various internal and external factors that might eventually trigger or lead to the disloyal act of an insider
- Opportunity, which depends on how easy it is for an insider to commit an attack. For example, insiders with more access rights or a system with more vulnerabilities would increase the opportunity for attack.

A recent systematic literature review by Gheyas and Abdallah (2016) categorized studies related to insider threat mitigation based on these three dimensions (or combination of dimensions). They concluded that the vast majority of studies falls into the category of opportunity. Specifically, about two-thirds of the studies examined in the systematic review used opportunity scores as the key features for insider threat detection and prediction. Drawing on an extensive range of sources, Gheyas and Abdallah (2016) summarized that most publications that concentrated on the opportunity dimension employed users' access rights and activities as indicators of opportunity. Users' access rights are defined by their system roles whereas the information on users' activities can be acquired from various types of log files such as database logs, web server logs, and error logs that provide a simple and cost-effective implementation for real-time activity tracking.

Compared to assessing the level of opportunity, assessing the level of motivation can be more challenging (Colwill, 2009), not only because of practical concerns in performing direct psychological examinations (Brown, Greitzer, & Watkins, 2013; Greitzer et al., 2010; Kiser et al., 2010), but also in terms of the difficulty in quantifying, recording, and tracking motivation systematically. The present study aims to add to this research line by demonstrating the viability of monitoring insiders'

motivation using automated text analyses. In the following subsection, we review some representative works that have adopted a similar methodological approach as presented in this study. A more comprehensive survey of the large body of research in insider threat and the structural organization of existing works based on various criteria can be found in Azaria, Richardson, Kraus, and Subrahmanian (2014) and Gheyas and Abdallah (2016).

2.2. Assessing Insiders' Motivation

Numerous frameworks have been proposed and applied to tackle the problem of insider threat from the insiders' motivation dimension. Research suggests that the motivation of an insider can be assessed from various aspects including the predisposition to malicious behavior, mental disorder, personality, and emotional states (Gheyas & Abdallah, 2016). Among these aspects, psychological and emotional vulnerabilities are often considered the pivotal factors contributing to insiders' motivation. For example, anger and disgruntlement have been frequently mentioned as indicating the motivation to perpetrate malicious conduct (Greitzer, Kangas, Noonan, Brown, & Ferryman, 2013; Ho et al., 2016; Shaw & Fischer, 2005). These negative psychological and emotional states can be shaped by interrelated factors coming from the inside and the outside, including one's personality, stress level, ability to cope with criticism, issues in personal life, and corporate factors, among others (Azaria et al., 2014).

Conceptually, many studies have suggested that assessing the emotional and psychological states of insiders can be useful for detecting and preventing malicious conduct. However, in terms of methodological development, there is still ample room for improvement in this research area. Axelrad, Sticha, Brdiczka, and Shen (2013) proposed the use of a Bayesian network for scoring insiders. The proposed model incorporated five categories of variables that measure occupational stress level and personal life stress level, personality variables, attitude and affect, history of social conflict, and so forth as indicators of an insider's degree of interest, which represents the relative risk of committing an attack. Based on empirical analysis of the collected data, the initial Bayesian network model was then adjusted to produce a predictive model of insider threat. Although their framework is conceptually appealing, the collection of data from a questionnaire poses some problems. The use of this data collection method is not uncommon in existing frameworks of insider threat. For instance, in Brdiczk et al. (2012), this method was used for psychological profiling in an insider prediction model. The development and validation of their model were carried out in the setting of a popular multi-player online game called World of Warcraft. Their framework combined structural anomaly detection of abnormal patterns in social and information networks with psychological profiling of the game characters to predict which characters would turn against their social groups in the game. For psychological profiling, the researchers made use of various sources including World of Warcraft census data, gamers' personality profiles from an online questionnaire, behavioral features based on gamers' activities in the game, simple analysis of game characters' names and their guild names, and the in-game social network of each gamer. In the model proposed by Kandias, Mylonas, Virvilis, Theoharidou, and Gritzalis (2010), the researchers also suggested the use of questionnaires to determine the stress level, predisposition, and user sophistication for psychological profiling of insiders. The shortcoming of this data collection method, however, is that it only allows periodical assessment, of which the frequency depends on practical factors like costs and cooperation from employees. Such a method might also be susceptible to self-reporting bias or other human-related biases. Contrastingly, automated text analysis methods facilitate continuous monitoring and reduce human biases.

In light of evidence that suggests a correlation between psychological and emotional factors and individuals' verbal (written or spoken) behavior (Pennebaker et al., 2001, 2003), there have been several attempts to predict or detect malicious insiders from numerous forms of textual data. One of the most prominent studies was undertaken by Greitzer et al. (2013). Based on the premises that psychosocial behavior is an indicator of insider threat and that these psychosocial factors are closely associated with word usage in spoken and written language, Greitzer et al. (2013) inferred that textual contents can be a valuable source for detecting insider threat through the identification of linguistic patterns pertaining to personality traits. Drawn upon a widely accepted standard for assessing personality traits, which is often referred to as the Big Five (McCrae, 2010), Greitzer et al. (2013) highlighted three personality traits—conscientiousness, neuroticism, and agreeableness—as the major factors that offer promise for prediction and detection of insider threat. The researchers applied text analysis to an email corpus that was injected with email samples of six known criminals. Their results showed that most of the known criminals were identified as outliers with high scores on neuroticism and low scores on conscientiousness and agreeableness.

In the same vein, Taylor et al. (2013) examined the changes in language usage in electronic communications when some team members decided to turn against the team. Their data

were collected from a simulated environment. The participants communicated through emails in a simulated workplace, whereby after the first stage of the study some participants were offered some incentives to start acting as malicious insiders in the team. From the text analysis of the collected emails, the researchers found that insiders showed several signs in their language usage: Compared to other co-workers, malicious insiders used more self-focused words, more negative language, and more words related to cognitive processes. Furthermore, the study also reported a deterioration in language similarity between insiders and other team members as the insiders became gradually estranged from the rest of the team over time.

Another study (Ho et al., 2016) applied linguistic analysis to conversational data collected from a multi-player gaming environment on the Google+ Hangout platform. The gaming environment simulated a betrayal scenario, in which a group member accepted an offer of incentives to betray the group. By comparing the within-group communications for those groups that did and did not have a deceptive insider (the control group), and before and after a member was compromised in a group, the study aimed to identify relevant linguistic cues such as negations, emotion-related words, words pertaining to cognitive processes, and so forth for revealing deceptive acts among the game players. The study reported that some subtle but identifiable patterns in group communications might represent an elevated risk of insider threat.

Collectively, all the studies described so far have provided evidence that text analysis can be a promising research direction towards understanding insider threat from a psychological perspective. However, these studies share a commonality; that is, all of them utilized Linguistic Inquiry and Word Count (LIWC), which is a text analysis program developed based on the works of Pennebaker and his colleagues (Pennebaker et al., 2001). Commonly used for psycholinguistic analysis, the LIWC program analyzes text by computing scores on word categories that provide insights into human social, cognitive, and affective dimensions. Although promising, existing studies remain narrow in focus, dealing mainly with text analysis using LIWC. An exception to this commonality is the framework proposed by Kandias, Stavrou, Bozovic, Mitrou, and Gritzalis (2013), that aimed to examine insiders' motivation by analyzing the content they generated and made public online. In their study, the researchers compared the performance of several machine learning techniques in classifying YouTube comments. The most accurate classifier was chosen as the model for detecting negative attitude towards authorities and enforcement of the law. Additionally, they also used a manually created, task-specific dictionary to perform classification by keyword matching. The framework proposed in the present study is similar to Kandias et al.'s framework in terms of the use of both statistical and dictionary-based methods for text analysis. However, while their framework only focused on identifying negative attitude, our framework covers a wider range of methods for a more comprehensive view that reveals the multiple facets of textual data using sentiment analysis, emotion analysis, and topic modeling.

3. METHOD

3.1. Proposed Framework

Fig. 1 shows the unified psycholinguistic framework proposed in the present study. This framework consolidates multiple

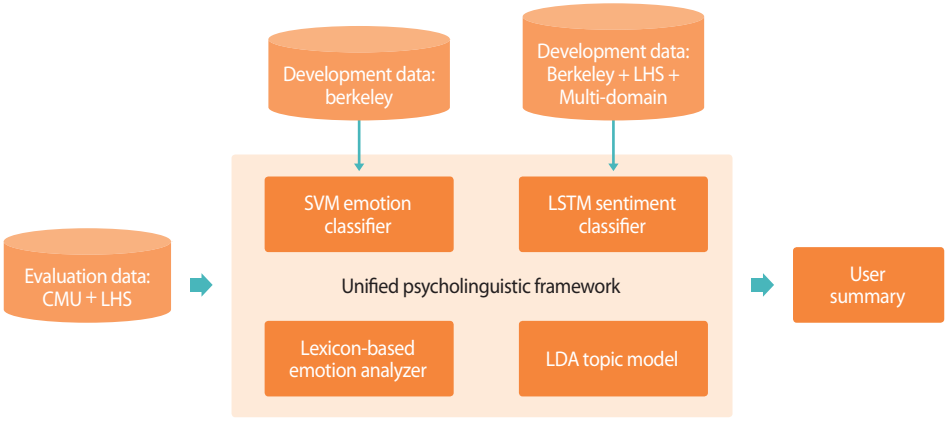


Fig. 1. The unified psycholinguistic framework. SVM, Support Vector Machine; LSTM, Long Short-Term Memory; LDA, Latent Dirichlet Allocation.

text analysis methods to generate comprehensive summaries of individuals' psychological states from their written texts. In relevance to our ultimate goal of supporting psychological assessments for prediction and detection of insider threat, we have adopted the following text analysis methods: sentiment analysis, emotion analysis, and topic modeling.

Sentiment analysis is a subfield of natural language processing that analyzes written or spoken language computationally to determine sentiment valences from textual contents. With regard to insider threat monitoring, sentiment analysis can provide an overview of whether an individual is a positive or negative person in general. In recent years, deep learning using neural networks has emerged as a powerful machine learning method for a diverse array of problems, including image processing, speech recognition, and various problems related to natural language processing. Like many machine learning methods, deep neural networks follow a data-driven approach to learn—from the training data—a function that best describes the mapping from the data to the output variable. One of the strengths of neural networks is their ability to represent a wide variety of mapping functions with very few constraints. As established by Hornik (1991) through theorem proving, with sufficient artificial neurons in the hidden layers, a multilayer neural network can be a universal approximator. The present study used Long Short Term Memory (LSTM) (Hochreiter & Schmidhuber, 1997), a recurrent neural network architecture that has been widely adopted for deep learning in many sentiment analysis tasks.

While sentiment analysis provides an overview of individuals' attitudes based on sentiment valences of texts, emotion analysis aims to give a more detailed view of individuals' affectual states such as angry, joyful, fear, sad, and so forth. The detection and classification of emotions have a wide range of applications, such as determining personality traits (Cherry, Mohammad, & De Bruijn, 2012) and detecting depression (Grijalva et al., 2015). With respect to the objective of the present study, we suggest that a closer look at individuals' emotions in addition to their sentiment valences can be useful for pinpointing potential threat. Specifically, emotion analysis can help to narrow down the list of possible suspects by targeting certain emotions. For instance, individuals associated with the anger emotion are probably more aggressive than individuals showing other negative emotions like sadness and fear. In the interest of exploring different approaches, we implemented two emotion analysis techniques in the present study: first, emotion classification with machine learning using Support Vector Machine (SVM) (Cortes & Vapnik, 1995); second, lexicon-based analysis using NRC emotion lexicon (Mohammad & Turney, 2013).

In addition to the analysis of sentiments and emotions, our proposed framework also includes topic modeling as one of its core components. Our topic model was built using Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), one of the most common topic modeling methods currently in use. The purpose of incorporating topic modeling into the framework is to give security analysts an overall picture of key topics penetrating the electronic communications under surveillance. Topic modeling can come in handy in this regard as it provides a convenient way to discover topical patterns statistically from the enormous volume of textual contents.

The implementation details of the text analyzers, as well as the data used for the development and evaluation of the proposed framework, are described in the following subsections.

3.2. Data

The following datasets were used for development and evaluation of the unified psycholinguistic framework:

- CMU Enron email dataset,[1] which is a collection of corporate emails of 150 users
- LHS dataset,[2] which consists of three types of text: love letters (L), hate emails (H), and suicide notes (S)
- UC Berkeley Enron email dataset,[3] which is a subset of an Enron email collection that contains 1,702 emotionally labeled emails
- Multi-domain review dataset,[4] which is a data corpus of positive and negative online reviews, ranging over 25 different topics including health, software, automotive, magazine, baby, beauty, and electronics, among others

3.2.1. Data for the Evaluation of the Framework

The evaluation of the framework was carried out on the first two datasets, i.e. the CMU Enron email dataset and the LHS dataset. Although the most ideal way to assess the effectiveness of the proposed framework is to evaluate it against ground truths of emotions, sentiments, and topics, such an approach would require large-scale manual labeling, which is time-consuming and costly. We thus resolved to verify the results generated by our framework using reference cases obtained from the LHS dataset. To this end, two synthetic users were created from each type of texts (i.e., L, H, and S) from the LHS dataset and were injected into the CMU Enron email dataset.

[1] https://www.cs.cmu.edu/~./enron

[2] http://saifmohammad.com/WebPages

[3] http://bailando.sims.berkeley.edu/enron_email.html

[4] https://www.cs.jhu.edu/~mdredze/datasets/sentiment

For instance, the 331 love letters from the LHS dataset were split into two sets consisting of 166 and 165 documents which constitute the 'emails' sent by synthetic users '_love1' and '_love2' respectively. Likewise, synthetic users '_hate1', '_hate2', '_suicide1', and '_suicide2' were created from hate emails and suicide notes.

Prior to text analysis, all emails in the CMU dataset were cleaned to remove email headers and forwarded texts. This step is essential to ensure that we only analyzed those emails that were sent—as opposed to received or forwarded—by the Enron users to understand the users' behavior from their written texts. This cleaning step was performed automatically by a computer script, followed by a manual examination of randomly selected emails to make sure that most emails were reasonably clean. This preprocessing step was carried out only on the CMU Enron dataset; the LHS dataset required no cleaning and was used in its original form. Altogether, the CMU Enron dataset and the LHS dataset resulted in a collection of texts contributed by 156 individuals.

3.2.2. Data for the Development of the Text Analyzers

Our unified psycholinguistic framework made use of both supervised and unsupervised methods for text analysis. The lexicon-based emotion analyzer and the topic model were implemented using unsupervised methods whereas the SVM emotion classifiers and the LSTM sentiment classifier were built via supervised learning, which entailed the use of manually labeled data in the development phase.

For the development of the SVM classifiers, the UC Berkeley Enron email dataset was used for training and testing the classification models. Although this dataset contains 19 emotion labels, only a subset of the labels is relevant to the goal of the present study. From Plutchik's model of emotions (Plutchik, 1982), we chose two emotions—anger and joy—for which classification models were built. As noted in many studies (Greitzer et al., 2013; Ho et al., 2016; Shaw & Fischer, 2005), the manifestation of the anger emotion in verbal communications can be a sign of psychological stress and dissatisfactions; it is thus not surprising that this emotion is often linked to the elevated risk of insider threat. The joy emotion was chosen as a contrasting emotion to anger because the low scores of joy emotion can somehow serve as supplementary evidence of negativity in individuals. Table 1 shows the mapping of emotions from the Berkeley dataset to the two emotions we are interested in. In addition to coalescing labeled emails into these main categories, we also replaced the emotion labels at the email-level with labels at the paragraph-level because quite often, lengthy emails that have been labeled with certain emotions only contain a few paragraphs pertaining to those emotions. It is thus reasonable to carry out emotion classification at the paragraph-level instead of email-level. To this end, the emails were automatically split into paragraphs at the occurrences of ending punctuation marks and empty lines, and the paragraphs that express the labeled emotions were identified manually. After email headers were removed from the data, the resulting paragraph-level dataset contains 37,684 instances, with 81 instances and 101 instances identified as pertaining to anger and joy, respectively.

Generally speaking, deep learning techniques like LSTM usually take a fairly large amount of data to achieve satisfactory performance. Therefore, for the development of the LSTM sentiment classifier, we combined the following: the UC Berkeley Enron email dataset, the LHS dataset, and the multi-domain review dataset. Synthetic users from the LHS dataset were first injected into the Berkeley dataset to obtain a bigger corpus. We assumed that all documents generated by '_love1' and '_love2' are positive whereas the documents generated by other synthetic users are negative. However, after combining the Berkeley Enron email dataset and the LHS dataset, the resulting dataset was rather imbalanced and inadequate for an optimum classification task. We thus further increased the size of the corpus with the multi-domain review dataset. Eventually, a balanced training dataset of 4,530 documents was obtained with 1,510 documents for each sentiment class. Unlike the classification of emotions, the classification of sentiments was implemented at the email-level instead of paragraph-level. The 19 emotions in the Berkeley dataset were mapped into positive, negative, and neutral sentiments as shown in Table 2.

Table 1. Mapping of emotions from the Berkeley dataset to three emotion categories

Emotion labels in the Berkeley dataset	Classified as
Anger / agitation	Anger
Jubilation and triumph / gloating	Joy
Humour, camaraderie, admiration, gratitude, friendship / affection, sarcasm, secrecy / confidentiality, concern, competitiveness / aggressiveness, pride, shame, hope / anticipation, dislike / scorn, worry / anxiety, sadness / despair, and sympathy / support	None (i.e., classified as no emotion)

3.3. Text Analyzers

In the interest of providing a more comprehensive view for monitoring users' emotional and psychological states, we have adopted a multi-faceted approach by including diverse types of text analyzers in our framework. The SVM and lexicon-based emotion analyzers help to identify individuals showing an exceptionally high level of anger emotion or an unusually low level of joy emotion; the LSTM sentiment classifier provides a view of individuals' positivity and negativity in general. The topic model aims to shed some light on the key topics around which the communications revolve.

3.3.1. SVM Emotion Classifiers

Using the paragraph-level Enron emails, we built SVM classifiers for binary classification of anger and joy such that each classifier was responsible for classifying every email paragraph as binary 1 or 0 based on the presence or absence of the emotions. Specifically, the anger classifier would classify a paragraph as presence (binary 1) if anger were detected in the paragraph, and absence (binary 0) if the paragraph showed no sign of anger. Likewise for the joy classifier. Our classification models used linear kernels with the cost of misclassification C = 0.1.

Building an SVM classifier via supervised learning entails finding the optimal decision boundary to separate instances of one class from another. In the SVM algorithm, this optimal decision boundary is the hyperplane that has the largest distance to the closest points of all classes. The performance of an SVM classifier relies heavily on the features that constitute the multi-dimensional feature space where the search for the best fitting hyperplane takes place. The present study made use of the WEKA package contributed by Mohammad and Bravo-Marquez (2017) to generate the following features for emotion classification:

- Word and character n-grams
- Negations: adding prefixes to words occurring in negated contexts. For instance, 'I do not like you' becomes 'I do not NEG-like NEG-you.'
- Part-of-speech tags: creating a vector space model from the sequence of part-of-speech tags
- Brown clusters: mapping words to Brown word clusters to create a low-dimensional vector space model
- Lexicon features: generating lexicon-related features using various lexicons including MPQA, Bing Liu's lexicon, AFINN, NRC Word-Emotion Association Lexicon, and so forth
- Positive and negative sentiment strengths: generating strengths of sentiments using SentiStrength

With the Enron paragraph-level emails, a major hindrance we faced in emotion classification was the huge discrepancy between the numbers of class 1 and class 0 instances. In other words, the dataset was highly imbalanced. This actually projects a realistic picture of the real-world data: In general, most emails are non-emotion-related and can be regarded as 'normal' emails, and the mission of the classifiers is to detect the tiny portion of 'abnormal' emails. Under such circumstances, most classifiers tend to bias towards the majority class, resulting in an extremely low (close to zero) accuracy in identifying instances of the minority class.

To tackle this problem, we applied under-sampling and over-sampling to reduce the gap between the majority class and the minority class. Under-sampling was first applied using random resampling to shrink the majority class to 30 times the size of the minority class. This was followed by the over-sampling procedure that uses Synthetic Minority Over-sampling Technique (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) to generate synthetic

Table 2. Mapping of emotions from the Berkeley dataset to the three sentiment classes

Sentiment labels	Emotion labels in the Berkeley dataset
Positive	Jubilation, hope / anticipation, humor, camaraderie, admiration, gratitude, friendship / affection, and sympathy / support
Negative	Worry / anxiety, concern, competitiveness / aggressiveness, triumph / gloating, pride, anger / agitation, sadness / despair, shame, and dislike / scorn
Neutral	Sarcasm and secrecy / confidentiality

Table 3. Numbers of instances in both classes before and after over-sampling and under-sampling

Emotion	Before over-sampling and under-sampling		After over-sampling and under-sampling	
	Presence	Absence	Presence	Absence
Anger	81	37,603	567	2,430
Joy	101	37,583	707	3,030

samples for the minority class, resulting in an expanded size of seven times the original size of the minority class. Note that the parameters that specified the scales by which these two classes were under-sampled or over-sampled were chosen by experiments. The original sizes of both classes and their sizes after over-sampling and under-sampling are given in Table 3.

3.3.2. LSTM Sentiment Classifier

We built the LSTM sentiment classifier using Keras, a high-level Python library that runs on top of Theano and TensorFlow to simplify the development of deep learning models. The network topology and the model parameters are described below.

- *Input layer*. Like other neural networks, an LSTM network requires numerical inputs. Therefore, text data need to be converted into numbers using word embedding—a text representation technique that maps discrete words into real-valued vectors. We chose to represent each word as a 128-dimensional vector, and the maximum length of each document in the dataset was capped at 50 words. In specific, the 4,530 documents in our dataset were converted to a set of 128 × 50 matrices.
- *Hidden layer*. The network's hidden layer contains 128 memory units. This layer takes as input the matrices generated by the word embedding representation procedure.
- *Output layer*. To tackle the three-class sentiment classification problem, the output layer of the network was designed as a dense (i.e., fully-connected) layer with three neurons and a softmax activation function to predict sentiment valences.

The network was trained for 10 epochs with a batch size of 32. Additionally, to reduce overfitting, we applied the dropout method to skip activation and weight updates for the inputs and recurrent connections at a probability of 0.2.

3.3.3. Lexicon-Based Emotion Analyzer

Lexical resources have been the key instrument in the analysis of sentiments and emotions. They provide scores, either discrete or continuous, for words and phrases that are salient indicators of sentiments and emotions. These resources can be utilized in many ways: Some studies used lexical resources as part of the rule-based approach while others incorporated lexicon-related features into the machine learning approach. But according to Mohammad (2015), the vast majority of works in emotion analysis have employed the statistical machine learning approach. One of the major obstacles in using the machine learning approach is the paucity of labeled data. As far as we know, large amounts of labeled data are only available from tweets, for which emoticons, emoji, and hashtag words such as #anger and #sadness can be used as emotion labels to produce pseudo-labeled data (Mohammad, 2012). Due to the limited amount of labeled data for emotion analysis of emails, in addition to the more widely used SVM emotion classification, we also implemented another emotion analyzer which relies merely on Mohammad and Turney's (2013) NRC emotion lexicon (version 0.92) to acquire emotion scores for the analyzed texts.

The NRC emotion lexicon provides binary values that indicate the presence or absence of Plutchik's eight emotions (Plutchik, 1982). Table 4 shows the binary values assigned to two examples of entries, 'abandonment' and 'helpful'. Using this lexicon, we followed the steps below to analyze anger and joy in the evaluation data:

- Lemmatization was first performed to preprocess the emails.
- Using the lemmatized texts, we looked up the NRC emotion lexicon to obtain a sum of scores for each emotion and for every email.
- To facilitate comparisons between individuals, we computed the final scores of anger and joy for every individual. Each final score was obtained by averaging the individual's overall score from all emails over the total number of emails written by the individual.

3.3.4. LDA Topic Model

One of the reasons for the emergence of topic modeling as a prevalent instrument for text analysis is the availability of many easy-to-use packages. In the present study, we used McCallum's topic modeling toolkit, MALLET, which provides a fast implementation of LDA (Blei et al., 2003), a method widely used for topic modeling and information summarization.

In the procedure of discovering latent topics from textual

Table 4. Examples of entries from the NRC emotion lexicon

Emotion	Abandonment	Helpful
Anger	1	0
Anticipation	0	0
Disgust	0	0
Fear	1	0
Joy	0	1
Sadness	1	0
Surprise	1	0
Trust	0	1

contents, LDA loops through all words in the text collection and assigns these words to the most probable topics. The procedure starts with a random assignment of topics and then rectifies the assignment over a large number of iterations until an optimal state is reached. In general, topic modeling considers a topic as a cluster of words that occur in some statistically meaningful ways. Given a text collection, the primary output of topic modeling is a list of keyword clusters pertaining to K topics, where K is a predetermined number that specifies how many topics are to be returned.

The present study applied LDA to extract 50 keyword clusters from all emails in the evaluation data. Labels were manually assigned to the 50 latent topics based on topic keywords and the most representative emails of each topic, i.e., emails with substantial contents related to the topics. Following the identification of these 50 most commonly discussed topics in our data, we then obtained the distribution of these key topics for each individual to perform cross-comparisons.

4. RESULTS AND DISCUSSION

Evaluations were carried out in two stages. At the first stage, text analyzers were evaluated individually to ensure that they are performing at the state-of-the-art level. At the second stage, the unified psycholinguistic framework as a whole was qualitatively assessed by its effectiveness in identifying potentially adversarial insiders through the use of various text analysis methods.

4.1. Evaluation of Individual Text Analyzers

This evaluation stage only applies to the SVM emotion classifiers and the LSTM sentiment classifier. Since these text analyzers were built via supervised learning, the learned classification models can be validated on the labeled development data. The precision (P), recall (R), and F-score (F) of emotion classification and sentiment classification are given by the equations below:

$$P = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalsePositive}} \quad (1)$$

$$R = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (2)$$

$$F = \frac{2PR}{P + R} \quad (3)$$

Since the evaluation data consists of highly imbalanced classes, the weighted averages of precision, recall, and F-score are used as the overall performance measures (Equations 4-6). For each of the k classes, the precision, recall, and F-score of the class are weighted by the number of instances in the class, and N is the total number of instances.

$$\text{WeightedAverage(P)} = \frac{\sum_{i=1}^{k} n_i P_i}{N} \quad (4)$$

$$\text{WeightedAverage(R)} = \frac{\sum_{i=1}^{k} n_i R_i}{N} \quad (5)$$

$$\text{WeightedAverage(F)} = \frac{\sum_{i=1}^{k} n_i F_i}{N} \quad (6)$$

The SVM emotion classifiers were validated with 3-fold cross-validation. Although 10-fold cross-validation is more commonly used for model validation, 3-fold cross-validation seems to be a better option in this case considering the number of class 1 instances of each emotion. In other words, it is unlikely that every partition would contain a sufficient number of class 1 instances if 10-fold cross-validation were to be used.

The validation was carried out in two experimental settings. In both settings, the classifier was trained with two-thirds of the over-sampled and under-sampled data in each round of the 3-fold cross-validation. Note that although the over-sampled minority class was used for training, the synthetic instances generated by Synthetic Minority Over-sampling Technique were excluded from the test data. In other words, the only difference between the two settings is the size of the majority class in the test data: While the first setting only tested the under-sampled majority class, the second setting tested all instances of the majority class. In the classification of highly imbalanced data, F-scores of the majority class (i.e., class 0) are usually beyond satisfactory. Therefore, our evaluation of the classifiers focuses mainly on the results obtained for the minority class (i.e., class 1), although the results for both classes are presented in Table 5 to give a complete picture of the classifiers' performance. Unless otherwise specified, the following discussion on the classifiers' performance refers to the minority class.

Overall, the results obtained with the under-sampled majority class are considered comparable to those demonstrated in existing studies (Mohammad, 2012; Mohammad, Zhu, Kiritchenko, & Martin, 2015). Nevertheless, when the emotion classifiers were validated with the full-sized majority class, the F-scores of both classifiers decreased tremendously. From the confusion matrices presented in Table 5, it can be seen that the

Table 5. Emotion classification results with 3-fold cross-validation

Emotion	Class	Classified as		Precision	Recall	F-score
		0	1			
With the under-sampled majority class						
Anger	0	2,423	7	0.985	0.997	0.991
	1	36	45	0.865	0.556	0.677
	Weighted average	-	-	0.981	0.983	0.981
Joy	0	3,015	15	0.987	0.995	0.991
	1	41	60	0.800	0.594	0.682
	Weighted average	-	-	0.981	0.982	0.981
With the full-sized majority class						
Anger	0	37,433	170	0.999	0.995	0.997
	1	33	48	0.220	0.593	0.321
	Weighted average	-	-	0.997	0.995	0.996
Joy	0	37,368	215	0.999	0.994	0.997
	1	39	62	0.224	0.614	0.328
	Weighted average	-	-	0.997	0.993	0.995

Table 6. Sentiment classification results with 10-fold cross-validation

Class	Classified as			Precision	Recall	F-score
	Neutral	Positive	Negative			
Neutral	1,374	63	73	0.896	0.910	0.903
Positive	81	1,111	318	0.752	0.736	0.744
Negative	78	303	1,129	0.743	0.748	0.745

degradation in F-scores was mainly caused by the increase in the number of false positives (i.e., false classification of class 0 as class 1). This is, however, not surprising in applications that aim to detect anomalous activities, which are rare and abnormal activities that may have serious consequences when not revealed. Apparently, the detection of potential insider threats also falls into this category of applications. Since the consequence of missing any potentially endangering individuals is detrimental, it is often necessary to have a 'skeptical' classifier to minimize the possibility of missing any suspects, even when it comes at the cost of higher false positive rate.

Table 6 shows the confusion matrix, precisions, recalls, and F-scores generated by the LSTM sentiment classifier with 10-fold cross-validation. The F-scores achieved on the prediction of the positive class (74.4%) and the negative class (74.5%) were considerably lower than the F-score achieved on the neutral class (90.3%). Nevertheless, the average F-score of the classifier (79.7%) still falls within an acceptable performance range for document-level sentiment analysis.

4.2. Evaluation of the Communications Using the Proposed Framework

Due to the inclusion of reference cases from the LHS dataset, at the time of evaluation we already had the prior knowledge that at least six of the 156 individuals in the evaluation data were affectively charged: two of them ('_love1' and '_love2') in a very positive way; four of them ('_hate1', '_hate2', '_suicide1', and '_suicide2') in a very negative way. For the proposed framework to be useful in identifying insider threat, it should be able to provide summaries of the users' psychological states so that informed decisions can be made to list the hate and suicide users as potentially malicious insiders and the love users as people who are not likely to involve in any adversarial acts. To demonstrate how this goal can be achieved by the proposed framework, we adopted an outlier detection method to select a small number of anomalous individuals from the evaluation data which comprises written texts contributed by 156 individuals. We then examined the cases of these anomalous individuals to assess the credence of the results produced by our framework.

The goal of outlier detection is to identify observations that deviate from common patterns and other observations in the collected data. Arriving at the conclusion that certain observations should be categorized as outliers is a highly subjective exercise. In the present study, we used interquartile range (IQR) and established the outlier range as 3 × IQR to identify outliers based on emotion scores obtained from the lexicon-based emotion analysis method described in Section 3.3.3. The outliers suggested by IQR are presented in Table 7. The scores shown next to the users' names indicate the average number of emotion words per email. For instance, '_hate2'—who tops other users at the anger emotion ranking—used an average of 5.73 anger words per email.

Judging from the rankings of the six reference cases, it seems the lexicon-based emotion analysis was able to generate reasonable scores for outlier detection. In particular, the love users are ranked on top of other users under the joy emotion whereas the hate and suicide users have higher rankings under the anger emotion. Another interesting finding is that the six reference cases always come before the real Enron users. This observation is likely to be related to the language and communication styles used in different forms of written communications: Compared to the more personal writing styles in love letters, hate mails, and suicide notes, business and professional writing in a workplace environment often uses a more formal tone and subtler expressions of personal emotion.

From the anomalous users shortlisted by the outlier detection method, we chose three individuals to zoom in into some user scenarios that would demonstrate the usage of the unified psycholinguistic framework and support the viability of predicting and detecting potential insider threats with psycholinguistic analysis. The following users were chosen for this purpose:

Table 7. Outliers detected using interquartile range on lexicon-based emotion scores

	Anger	Joy
1	_hate2 (5.73)	_love1 (12.54)
2	_suicide2 (4.91)	_love2 (11.14)
3	_hate1 (4.51)	_suicide2 (7.64)
4	_suicide1 (4.1)	_suicide1 (7.1)
5	_love2 (1.49)	_hate2 (6.44)
6	_love1 (1.38)	_hate1 (4.17)
7	_enron1 (1.27)	
8	_enron2 (1.09)	
9	_enron3 (1.04)	
10	_enron4 (1.03)	
11	_enron5 (1.01)	

- '_love2' and '_hate2'. These synthetic users serve the purpose of quick verification for the proposed framework.
- '_enron1'. This user is the Enron employee that scored highest under the anger emotion.

The user scenarios of the three users are presented with the following graphs and charts that visualize the results generated by all text analyzers in the proposed framework:

- The lexicon-based emotion timeline shows the emotion scores obtained from a simple count of emotion words per email. The scores were normalized to the range of [min, max], where min and max are the minimum and maximum count of anger-related or joy-related words—depending on which emotion is analyzed—across the three users included in the user scenarios.
- The sentiment classification timeline and proportion chart visualize the emails' class labels (positive / *pos*, neutral / *neu*, and negative / *neg*) predicted by the LSTM sentiment classifier.
- The emotion classification timeline and proportion chart visualize the emails' class labels (1 for presence, 0 for absence) predicted by the SVM emotion classifiers. Since the classifiers were trained at the paragraph-level instead of document-level, the predictions were first carried out at paragraph-level but a final class label was obtained for each email using a logical OR function, which assigned class label 1 to an email if any of its paragraphs was predicted as 1 in the classification.
- The topic distribution chart shows the counts of users' emails pertaining to the 50 topics extracted by the LDA topic model.

In the scenarios of the two synthetic users, what stands out the most is that the results produced by all text analyzers seem to agree with each other. Although it would be overstating matters to claim that they are highly similar, we can still conclude that the emotion scores generated by the lexicon-based emotion analyzer and the predictions made by the LSTM sentiment classifier and the SVM emotion classifiers show convincing similarity to a certain extent. For example, in the scenario of the love user, the lexicon-based emotion timeline (Fig. 2) depicts that the number of joyful words in the user's texts clearly surpasses the number of angry words. In agreement with this outcome, the predictions provided by the sentiment classifier and the emotion classifiers show that a large portion of the _love2 user emails are positive (Fig. 3) and joyful (Fig. 4). Likewise, the _hate2 user's emotion scores obtained from lexicon-based emotion analysis (Fig. 5) and predictions

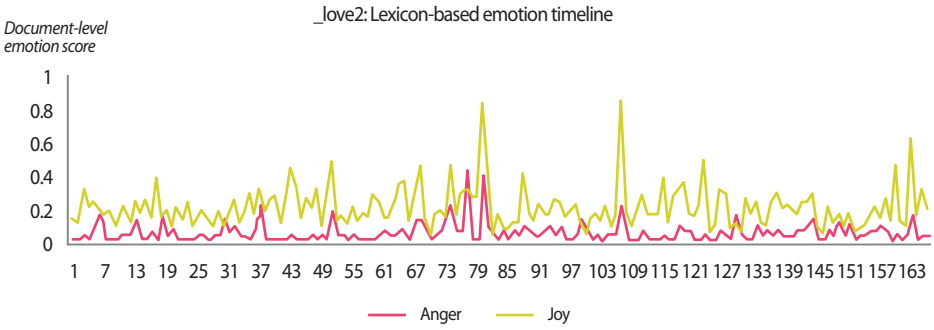


Fig. 2. Lexicon-based emotion timeline for synthetic user '_love2.'

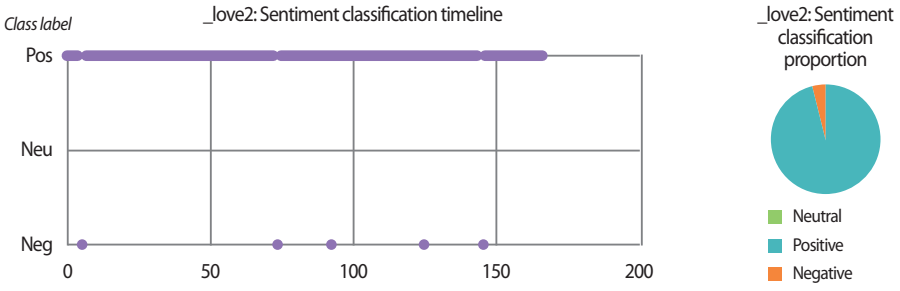


Fig. 3. Sentiment classification timeline and proportion for synthetic user '_love2.'

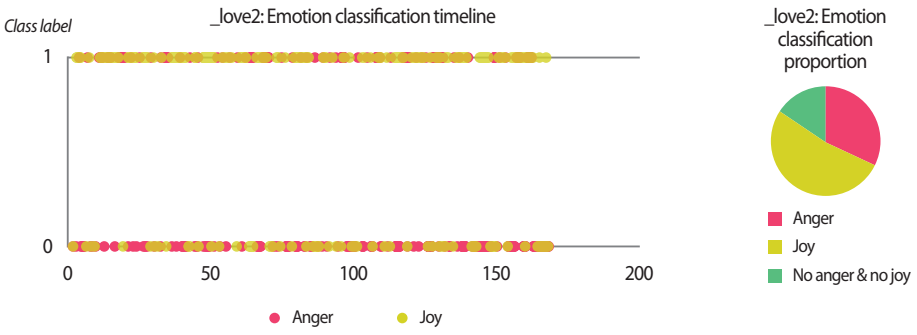


Fig. 4. Emotion classification timeline and proportion for synthetic user '_love2.'

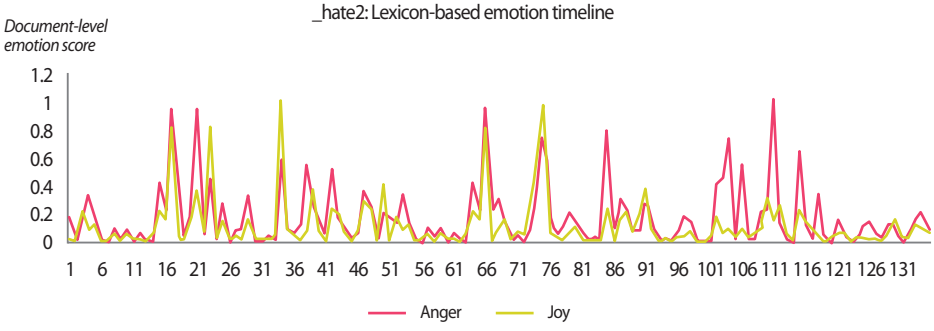


Fig. 5. Lexicon-based emotion timeline for synthetic user '_hate2.'

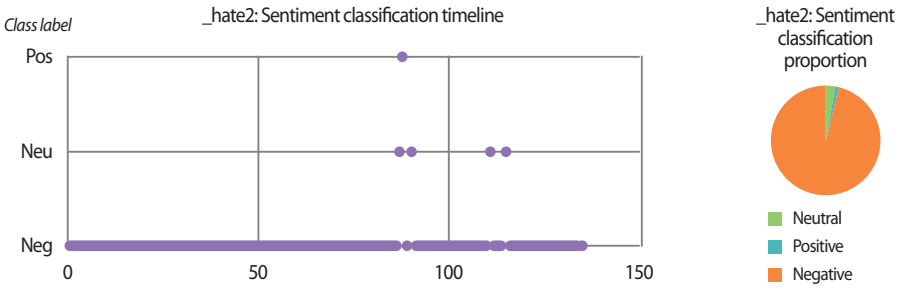


Fig. 6. Sentiment classification timeline and proportion for synthetic user '_hate2.'

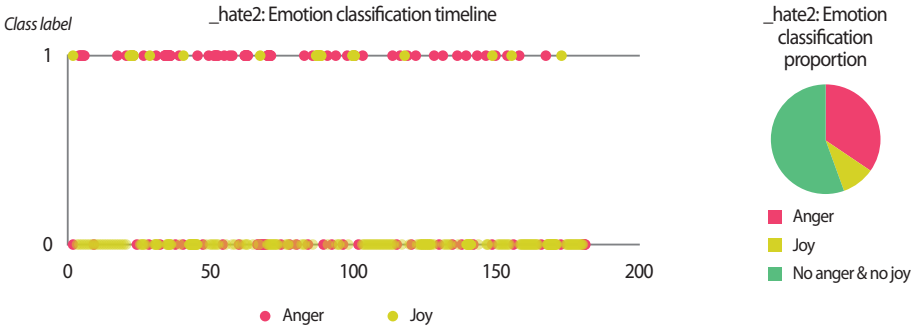


Fig. 7. Emotion classification timeline and proportion for synthetic user '_hate2.'

generated by the classifiers (Figs. 6 and 7) provide a lens into the anger and negativity in the user.

Turning now to the user scenario of the Enron user '_enron1,' we noticed that the three text analyzers for sentiment analysis and emotion analysis produced contradictory results. While the lexicon-based emotion analyzer and the LSTM sentiment classifier detected more emails that showed negative sentiments and emotions, the SVM emotion classifiers reported that the _enron1 user's emails did not contain any angry or hateful texts. Since this Enron user only produced 11 emails in his sent folder, we were able to label all emails manually to provide ground truths for verification (Table 8).

By comparing the results in Figs. 9 and 10 to the manual labels in Table 8, it can be seen that neither the sentiment classifier nor the emotion classifiers produced accurate predictions, although the predictions obtained by the sentiment classifier are slightly more accurate than the predictions generated by the emotion classifiers. However, from a closer inspection of the lexicon-based emotion timeline shown in Fig. 8, we found that there are two spikes in the anger emotion timeline that matched the manual labels in Table 8. One spike occurred on 21/2/2001 and the other spike occurred on 28/6/2001. Based on the textual contents of the two emails sent by '_enron1' on 21/2/2001 and 28/6/2001 (Fig. 11), it seems the lexicon-based emotion analysis has revealed a remarkable potential for the detection of emotions in electronic communications.

Table 8. Manually labeled sentiments and emotions for emails sent by '_enron1'

Email date	Sentiment	Anger	Joy
21/2/2001 23:09	-1	1	0
26/2/2001 5:34	0	0	0
28/2/2001 3:22	0	0	0
9/4/2001 10:06	1	0	0
1/5/2001 11:11	0	0	0
10/5/2001 1:56	0	0	0
11/5/2001 6:18	0	0	0
11/5/2001 7:03	1	0	0
15/5/2001 0:13	0	0	0
28/6/2001 1:25	-1	1	0
29/10/2001 16:38	0	0	0

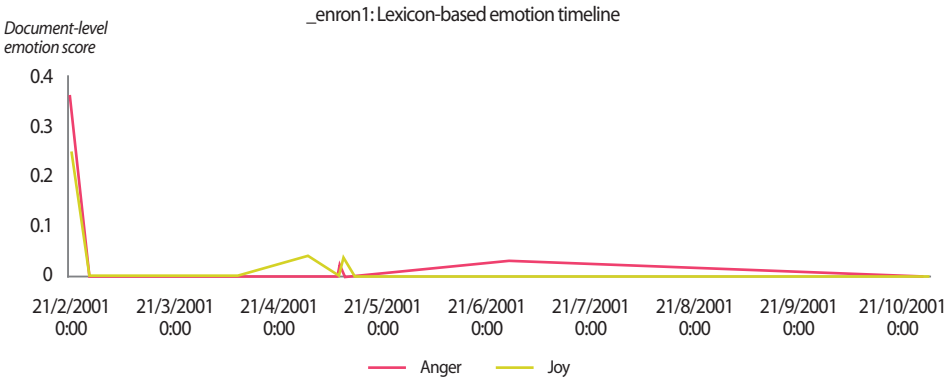


Fig. 8. Lexicon-based emotion timeline for Enron user '_enron1.'

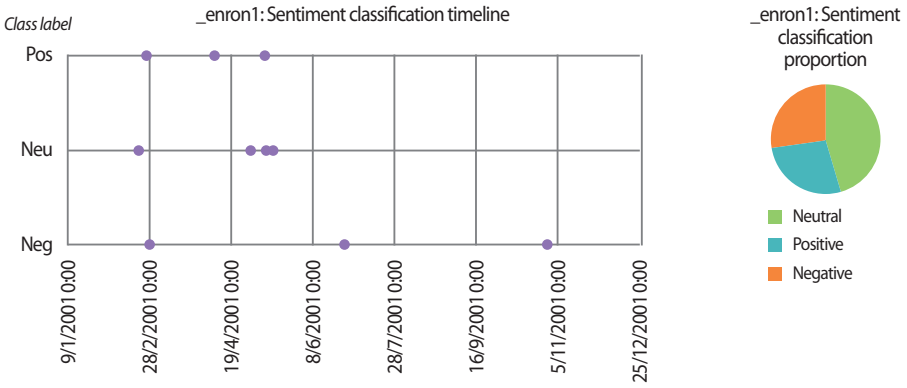


Fig. 9. Sentiment classification timeline and proportion for Enron user '_enron1.'

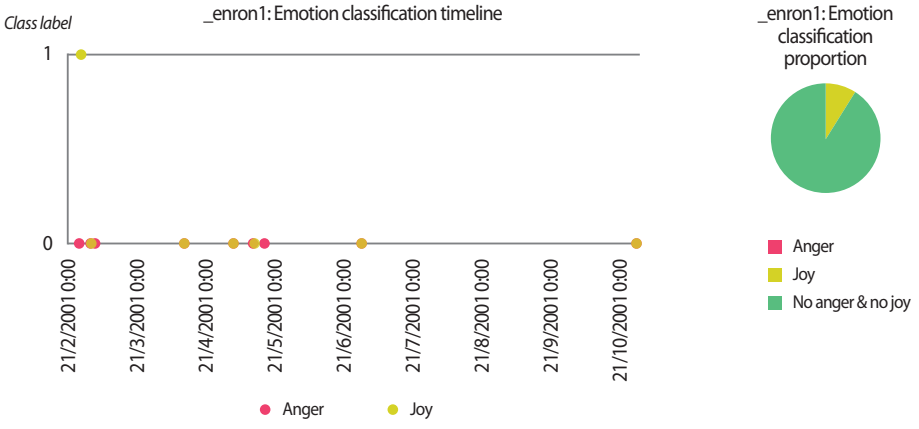


Fig. 10. Emotion classification timeline and proportion for Enron user '_enron1.'

Email on 21/2/2001 23:09

The Following are our recommended changes to the agreement(I would also like to discuss the waiver of ***conflicts*** *section with you):*

.....

1. With respect to Haywood Power I, L.L.C. ("Haywood Power"), the ***abandonment*** *by the Tennessee Valley Authority ("TVA") of its proposed 320 MW expansion of its Haywood County, Tennessee facility ("Lagoon Creek"), which results in Haywood Power being allowed by TVA to interconnect into an existing 500kV open bus position in the Lagoon Creek substation. Achievement of this milestone shall be evidenced by the interconnection specifications set forth in an Interconnection Agreement between TVA and Haywood Power, and the milestone shall be deemed to be achieved upon execution of said Interconnection Agreement.*

2. With respect to Haywood Power, the decision by TVA to ***eliminate*** *the Network Upgrade related to the reactive power requirements as set forth in Haywood Power's System Impact Study, presently estimated at a total cost of $5 million. Achievement of this milestone shall be evidenced by the Network Upgrade requirements set forth in an Interconnection Agreement between TVA and Haywood Power, and the milestone shall be deemed to be fully achieved upon execution of said Interconnection Agreement to the extent such Network Upgrade cost is* ***eliminated****. To the extent that such Network Upgrade cost is less than $5 million, but greater than zero, a pro rata portion of the $175,000 project fee shall be paid.*

3. With respect to Calvert City Power I, L.L.C. ("Calvert"), the ***elimination*** *of TVA's present requirement for*

.....

Email on 28/6/2001 1:25

If this is who(Governor Davis) DGA and NDN chooses to support despite the fact that his ***lack*** *of leadership has imposed incalculble* ***pain*** *and costs on California, please* ***remove*** *me from your e-mail and any other lists from here forward. Thank you.*

Fig. 11. Emails sent by '_enron1' on 21/2/2001 and 28/6/2001. Words that signify anger and negativity are underlined.

In addition to understanding individuals' psychological states from sentiment analysis and emotion analysis, we also extracted 50 key topics from the data corpus using the LDA topic model. The list of keywords composing the 50 topics is given in Appendix A. For each user scenario, the topic distribution and the top three topics discussed by the user are presented in the topic distribution charts (Figs. 12-14). The top topics revealed that the emails written by '_enron1' were mainly business-related, covering topics on payments and charges, regulatory concerns, secretarial communications, and so forth. On the other hand, the synthetic users tend to discuss matters revolving around more personal themes like blessings and wishes, fun comments, and criticisms. Another interesting finding from topic modeling is that key topics also reflect the sentiments and emotions of users. For instance, the top topic of '_hate2' (i.e., criticisms and negative reactions) shows that this user had a tendency to criticize and react negatively. Likewise, the positive behavioral patterns of '_love2' can be easily spotted from the user's top topics. These results suggest that topic modeling can be a sensible supplementary technique for assessing individuals' psychological states from their verbal communications.

Taken together, the user scenarios presented so far demonstrated how the unified psycholinguistic framework can keep the false alarm rate at a manageable level without compromising the detection of potential insider threats. The dilemma has been addressed in two ways: First, as seen from the scenarios of the synthetic users, the uncertainty in the insider tracking process can be reduced considerably when multiple text analyzers agree with each other; second, as demonstrated by the scenario of the Enron user, multiple text analyzers might complement each other and produce contradictory results in some cases. This scenario can be taken as an indicator for invoking a follow-up investigation by a human analyst to minimize the risk of missed catches.

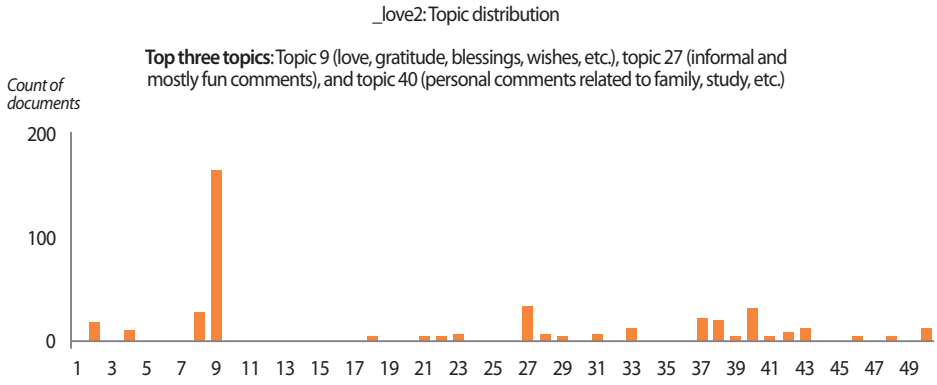


Fig. 12. Topic distribution and the top three topics for synthetic user '_love2.'

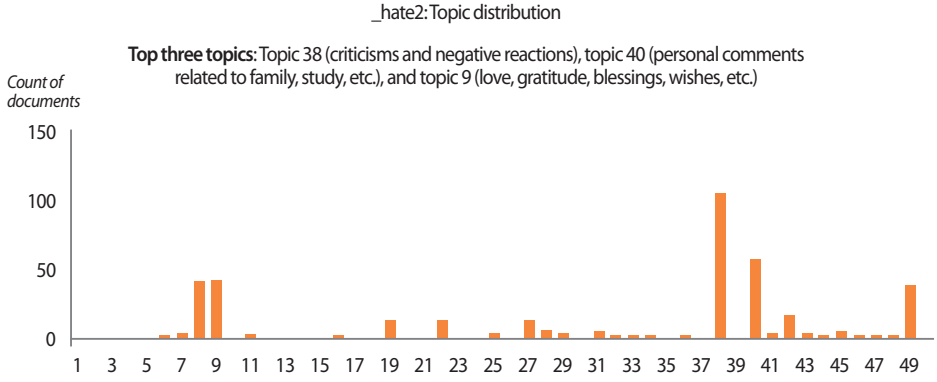


Fig. 13. Topic distribution and the top three topics for synthetic user '_hate2.'

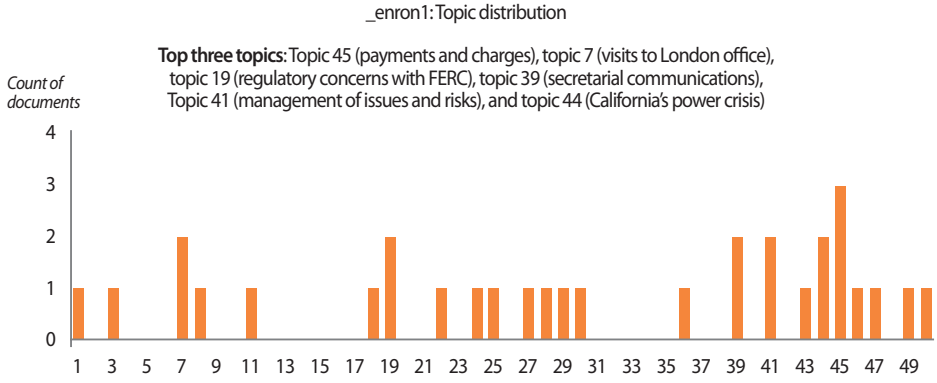


Fig. 14. Topic distribution and the top three topics for Enron user '_enron1.'

5. CONCLUSION

The present study was undertaken to predict and detect insider threat by monitoring electronic communications for identifying individuals with troubling psychological patterns. To that end, we combined several text analysis methods—lexicon-based emotion analysis, LSTM sentiment classification, SVM emotion classification, and LDA topic modeling—to form a unified psycholinguistic framework. This is the first study that examined the use of multiple text analysis methods for psycholinguistic assessment in insider threat mitigation. The user scenarios presented in this paper demonstrated how the issue of the trade-off between the risk of missed catches and the false alarm rate can be attenuated. Overall, the text analyzers in our framework achieved acceptable performance. Further improvement is possible but is limited by some known constraints, such as highly imbalanced classes and the paucity of labeled data. In terms of directions for future research, considerably more work will need to be done to overcome these constraints and to achieve better accuracy in sentiment classification and emotion classification. Another natural progression of this work is to carry out the evaluation of the framework on data containing real or simulated insider threat.

REFERENCES

Axelrad, E. T., Sticha, P. J., Brdiczka, O., & Shen, J. (2013). A Bayesian network model for predicting insider threats. In *Proceedings of the 2013 IEEE Security and Privacy Workshops* (pp. 82-89). Piscataway: IEEE.

Azaria, A., Richardson, A., Kraus, S., & Subrahmanian, V. S. (2014). Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data. In *IEEE Transactions on Computational Social Systems* (pp. 135-155). Piscataway: IEEE.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Brdiczka, O., Liu, J., Price, B., Shen, J., Patil, A., Chow, R., . . . Ducheneaut, N. (2012). Proactive insider threat detection through graph learning and psychological context. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy Workshops* (pp. 142-149). Piscataway: IEEE.

Brown, C. R., Greitzer, F. L., & Watkins, A. (2013). Toward the development of a psycholinguistic-based measure of insider threat risk focusing on core word categories used in social media. In *AMCIS 2013 Proceedings* (pp. 3596-3603). Atlanta: Association for Information Systems.

Brown, C. R., Watkins, A., & Greitzer, F. L. (2013). Predicting insider threat risks through linguistic analysis of electronic communication. In *Proceedings of the 46th Hawaii International Conference on System Sciences* (pp. 1849-1858). Piscataway: IEEE.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321-357.

Chen, Y., & Malin, B. (2011). Detection of anomalous insiders in collaborative environments via relational analysis of access logs. In *Proceedings of the First ACM Conference on Data and Application Security and Privacy* (pp. 63-74). New York: ACM.

Cherry, C., Mohammad, S. M., & De Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5(Suppl 1), 147-154.

Colwill, C. (2009). Human factors in information security: The insider threat–Who can you trust these days?. *Information Security Technical Report*, 14(4), 186-196.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.

Eberle, W., Graves, J., & Holder, L. (2010). Insider threat detection using a graph-based approach. *Journal of Applied Security Research*, 6(1), 32-81.

Gheyas, I. A., & Abdallah, A. E. (2016). Detection and prediction of insider threats to cyber security: A systematic literature review and meta-analysis. *Big Data Analytics*, 1(1), 6.

Greitzer, F. L., Frincke, D. A., & Zabriskie, M. (2010). Social/ethical issues in predictive insider threat monitoring. In M. J. Dark (Ed.), *Information assurance and security ethics in complex systems: Interdisciplinary perspectives* (pp. 1100-1129). Hershey: IGI Global.

Greitzer, F. L., Kangas, L. J., Noonan, C. F., Brown, C. R., & Ferryman, T. (2013). Psychosocial modeling of insider threat risk based on behavioral and word use analysis. *e-Service Journal*, 9(1), 106-138.

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., & Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological Bulletin*, 141(2), 261-310.

Ho, S. M., Hancock, J. T., Booth, C., Burmester, M., Liu, X., & Timmarajus, S. S. (2016). Demystifying insider threat: Language-action cues in group dynamics. In *Proceedings of the 49th Hawaii International Conference on System Sciences* (pp. 2729-2738). Piscataway: IEEE.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735-1780.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251-257.

Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., & Gritzalis, D. (2010). An insider threat prediction model. In S. Katsikas, J. Lopez, & M. Soriano (Eds.), *Lecture notes in computer science: Vol. 6264. Trust, privacy and security in digital business* (pp. 26-37). Berlin: Springer.

Kandias, M., Stavrou, V., Bozovic, N., Mitrou, L., & Gritzalis, D. (2013). Can we trust this user? Predicting insider's attitude via YouTube usage profiling. In *Proceedings of the 2013 IEEE 10th International Conference on Ubiquitous Intelligence and Computing and 2013 IEEE 10th International Conference on Autonomic and Trusted Computing* (pp. 347-354). Piscataway: IEEE.

Kiser, A. I., Porter, T., & Vequist, D. (2010). Employee monitoring and ethics: Can they co-exist?. *International Journal of Digital Literacy and Digital Competence*, 1(4), 30-45.

McCrae, R. R. (2010). The place of the FFM in personality psychology. *Psychological Inquiry*, 21(1), 57-64.

Mohammad, S. M. (2012). #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 246-255). Stroudsburg: Association for Computational Linguistics.

Mohammad, S. M. (2015). Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In H. L. Meiselman (Ed.), *Emotion measurement* (pp. 201-237). Duxford: Woodhead Publishing.

Mohammad, S. M., & Bravo-Marquez, F. (2017). Emotion intensities in tweets. In *Proceedings of the Sixth Joint Conference on Lexical and Computational Semantics* (pp. 65-77). Stroudsburg: Association for Computational Linguistics.

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3), 436-465.

Mohammad, S. M., Zhu, X., Kiritchenko, S., & Martin, J. (2015). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4), 480-499.

Myers, J., Grimaila, M. R., & Mills, R. F. (2009). Towards insider threat detection using web server logs. In *Proceedings of the 5th Annual Workshop on Cyber Security and Information Intelligence Research: Cyber Security and Information Intelligence Challenges and Strategies* (p. 54). New York: ACM.

Parker, D. B. (1998). *Fighting computer crime: A new framework for protecting information*. New York: Wiley.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2001). *Linguistic inquiry and word count: LIWC 2001*. Retrieved February 5, 2019 from http://www.depts.ttu.edu/psy/lusi/files/LIWCmanual.pdf.

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547-577.

Plutchik, R. (1982). A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5), 529-553.

Schultz, E. E. (2002). A framework for understanding and predicting insider attacks. *Computers & Security*, 21(6), 526-531.

Shaw, E. D., & Fischer, L. F. (2005). *Ten tales of betrayal: The threat to corporate infrastructure by information technology insiders analysis and observations*. Retrieved February 5, 2019 from http://www.dtic.mil/dtic/tr/fulltext/u2/a441293.pdf.

Taylor, P. J., Dando, C. J., Ormerod, T. C., Ball, L. J., Jenkins, M. C., Sandham, A., & Menacere, T. (2013). Detecting insider threats through language change. *Law and Human Behavior*, 37(4), 267-275.

Wood, B. (2000). An insider threat model for adversary simulation. In *Proceedings of the Workshop on Mitigating the Insider Threat to Information Systems* (pp. 41-48). Arlington: RAND.

APPENDIX A

Keywords for the 50 Topics Generated from Topic Modeling

Topic ID	Keywords
1	power project system plant gas fuel approved transmission development cost production process run load opportunity request gathering capacity proposal approval
2	hope great happy management game home tickets tonight fun nice night sat weekend guys party play christmas birthday stay hey
3	kay ge stuff don email ben working lisa lee docs ena sheila change llc closing emails reminder ready equipment advise
4	time free weeks couple feel wanted told lunch interested guys week meet called today forward taking thought place ago asked
5	mark enron greg mary david works delainey john taylor counsel ed andy liz tim general group fletch president whalley services
6	deal deals volume month jan ces sitara desk storage feb oct nov created booked volumes dth sale entered book changed
7	mail london message houston df office voice received check leave pls trip flight calendar hotel address number wed messages travel
8	people money told years big story made lot things find country worth talking didn making world times hard past running
9	love life day feel man time heart dear ve make world god true things mind happy baby met words kind
10	enron corp houston north america eb texas street smith legal debra perlingiere department sara shackleton fax phone gisb tx ph
11	business review employees global enron process ena performance prc employee rick year focus feedback commercial meetings level training committee management
12	phone fax sara st susan carol confirm cell clair lawyer send spoke ss confirms lawyers confirmation handling suzanne leslie reach
13	trading trade financial counterparty book products eol physical credit master trades online canada counterparties product swap power transactions books legal
14	meeting monday friday pm thursday afternoon tuesday attend wednesday morning meet tomorrow office schedule week scheduled set noon attending unable
15	call give jeff discuss ll conference tomorrow today folks number set chance df thoughts heather asap apologies answer srs voicemail
16	report desk data information open phillip position west positions spreadsheet reports update items run mat find format track summary var
17	attached comments draft agreement questions form version forward revised request discussed latest final document approval hesitate prepare proposed agreements attaching
18	sally office team week meeting plan operations group brent patti work key james memo meetings join working role review calgary
19	information ferc options option case provide policy specific terms part concerns aware regulatory made related confidential additional including concern reference
20	mid kate dec changed columbia pst deals deal ees broker morgan bpa avista epmi stephanie mc pget cob sp enpower
21	talk today chris yesterday ben tomorrow morning talked wanted didn robin matt hey lets brian joe chicago ya don working
22	make access put bill note group page set read line idea computer change work picture suggestions direct time link mind
23	day time date days june july december april march plan october end vacation august september january schedule november dates back
24	person left kevin check info message michael city eric gary rob tx pass jason east portland julie leave asked ckm
25	agreement credit language ena section party master isda guaranty transaction agreements parties termination contract paragraph assignment respect dated transactions law
26	gas daily index price basis el socal paso mmbtu east pg day pool volumes west hpl curve point ena flow
27	don ve didn guy thing remember ll back guys bad pretty doesn couple stuff thought half figure worry couldn finally
28	year program game team play big center ut university end top students early recruiting round school turn football national won
29	market price prices million year power term demand costs supply cost based long increase high percent buy summer sales cap
30	fyi jim file kim info tom michelle fine notes dg files update lynn questions linda jennifer coordinate harry alan fred
31	issues issue make order problem agree response understand problems decision point future clear discuss made credit end comment suggest line
32	contract rate capacity tw service contracts firm point delivery pipeline term df tariff ena rates order release fuel transportation points
33	good great hope sounds job things work hear glad pretty luck guys thought hey talk interesting summer care trip nice
34	send copy letter review sign signed documents print copies received executed sending lynn elizabeth marie original document attachment attached signature
35	john dave forward paul rick resume dan interest manager interested interview eric congratulations frank peter director robert follow michelle directly
36	buy mw bid short hours offer blackberry sale wireless handheld sell peak hour bill day real schedule power purchase show
37	work week time make back move lot start working good things hard long moving weather ready doesn rest busy happen
38	don people question sense makes answer fact make site read opinion doesn case find problem long situation isn wrong website
39	vince jeff presentation steve ken mr shirley assistant lay fyi skilling invite kaminski invitation stinson speak karen join sherri george
40	school work years family class part visit hope live dr remember care day children miss find write don parents pictures
41	group risk involved model project management support working provide current experience manage issues position area process structure large additional including
42	room house place bring car front church side black water small red hit stop white walk door parking boat put
43	ll back don ve heard haven today check email guess fine hear waiting chance wait touch password checked talk chat
44	california state power energy davis utilities edison utility electricity contracts commission puc customers governor dwr billion public plan pg bill
45	pay amount payment cash contract notice tax paid due account charge made period purchase event charges money receive fee days
46	facility test site unit prior permit units transwestern completed required submitted request issue air construction additional mexico system activities agency
47	mike list email contact send address bob add scott names distribution asked update stacey jay message david missing forwarded richard
48	number pl numbers questions change sheet put correct call problem find stan errol note today louise added cindy verify worksheet
49	enron company energy services business trading gas marketing markets stock companies natural capital resources corporation interest europe officer york board
50	night weekend dinner home saturday sunday town house dad tonight mom leaving leave kids fun friday coming friend austin plans

Call for Paper

Journal of Information Science Theory and Practice (JISTaP)

We would like to invite you to submit or recommend papers to **Journal of Information Science Theory and Practice** (JISTaP, eISSN: 2287-4577, pISSN: 2287-9099), a fast track peer-reviewed and no-fee open access academic journal published by Korea Institute of Science and Technology Information (KISTI), which is a government-funded research institute providing STI services to support high-tech R&D for researchers in Korea. JISTaP marks a transition from Journal of Information Management to an English-language international journal in the area of library and information science.

JISTaP aims at publishing original studies, review papers and brief communications on information science theory and practice. The journal provides an international forum for practical as well as theoretical research in the interdisciplinary areas of information science, such as information processing and management, knowledge organization, scholarly communication and bibliometrics.

We welcome materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. Topics covered by the journal include: information processing and management; information policy; library management; knowledge organization; metadata and classification; information seeking; information retrieval; information systems; scientific and technical information service; human-computer interaction; social media design; analytics; scholarly communication and bibliometrics. Above all, we encourage submissions of catalytic nature that explore the question of how theory can be applied to solve real world problems in the broad discipline of information science.

Co-Editors in Chief : Gary Marchionini & Dong-Geun Oh

Please click the "Online Submission" link in the JISTaP website (http://www.jistap.org), which will take you to a login/ account creation page. Please consult the "Author's Guide" page to prepare your manuscript according to the JISTaP manuscript guidelines.

Any question? Ji-Young Kim (managing editor) : jistap@kisti.re.kr

Information for Authors

The Journal of Information Science Theory and Practice (JISTaP), which is published quarterly by the Korea Institute of Science and Technology (KISTI), welcomes materials that reflect a wide range of perspectives and approaches on diverse areas of information science theory, application and practice. JISTaP is an open access journal run under the Open Access Policy. See the section on Open Access for detailed information on the Open Access Policy.

A. Originality and Copyright

All submissions must be original, unpublished, and not under consideration for publication elsewhere. Once an article is accepted for publication, all papers are accessible to all users at no cost. If used for other researches, its source should be indicated in an appropriate manner and the content can only be used for uncommercial purpose under Creative Commons license.

B. Peer Review

All submitted manuscripts undergo a single-blind peer review process in which the identities of the reviewers are withheld from the authors.

C. Manuscript Submission

Authors should submit their manuscripts online via Article Contribution Management System (ACOMS). Online submission facilitates processing and reviewing of submitted articles, thereby substantially shortening the paper lifecycle from submission to publication. After checking the manuscript's compliance to the Manuscript Guidelines, please follow the "Online Submission" hyperlink in the top navigation menu to begin the online manuscript submission process.

D. Open Access

With the KISTI's Open Access Policy, authors can choose open access and retain their copyright or opt for the normal publication process with a copyright transfer. If authors choose open access, their manuscripts become freely available to public under Creative Commons license. Open access articles are automatically archived in the KISTI's open access repository (KPubS, www.kpubs.org). If authors do not choose open access, access to their articles will be restricted to journal users.

E. Manuscript Guidelines

Manuscripts that do not adhere to the guidelines outlined below will be returned for correction. Please read the guidelines carefully and make sure the manuscript follows the guidelines as specified. We strongly recommend that authors download and use the manuscript template in preparing their submissions.

1. Page Layout :
All articles should be submitted in single column text on standard Letter Size paper (21.59 × 27.94 cm) with normal margins.

2. Length :
Manuscripts should normally be between 4,500 and 9,000 words (10 to 20 pages).

3. File Type :
Articles should be submitted in Microsoft Word format. To facilitate the manuscript preparation process and speed up the publication process, please use the manuscript template.

4. Text Style :
- Use a standard font (e.g., Times New Roman) no smaller than size 10.
- Use single line spacing for paragraphs.
- Use footnotes to provide additional information peripheral to the text. Footnotes to tables should be marked by superscript lowercase letters or asterisks.

5. Title Page :
The title page should start with a concise but descriptive title and the full names of authors along with their affiliations and contact information (i.e., postal and email addresses). An abstract of 150 to 250 words should appear below the title and authors, followed by keywords (4 to 6).

Author1
Affilliation, Postal Address. E-mail

Author2
Affilliation, Postal Address. E-mail

ABSTRACT
A brief summary (150-250 words) of the paper goes here.

Keywords : 4 to 6 Keywords, separated by commas.

6. Numbered Type :

1. INTRODUCTION
All articles should be submitted in single column text on standard letter size paper (21.59 × 27.94 cm) with normal margins[1 . Text should be in 11-point standard font (e.g., Times New Roman) with single line spacing.
[1 Normal margin dimensions are 3 cm from the top and 2.54 cm from the bottom and sides.

2. SECTIONS

The top-level section heading should be in 14-point bold all uppercase letters.

2.1. Subsection Heading 1

The first-level subsection heading should be in 12-point bold with the first letter of each word capitalized.

2.1.1. Subsection Heading 2

The second-level subsection heading should be in 11-point italic with the first letter of each word capitalized.

7. Figures and Tables :

All figures and tables should be placed at the end of the manuscript after the reference list. To note the placement of figures and tables in text, "Insert Table (or Figure) # here" should be inserted in appropriate places. Please use high resolution graphics whenever possible and make sure figures and tables can be easily resized and moved.

Figure

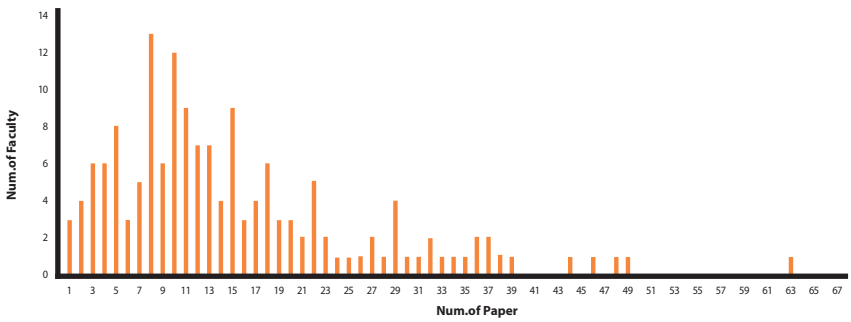


Fig. 1. Distribution of authors over publication count.

Table

Table 1. The title of table goes here

Study	Time period study	Data
Smith Wesson (1996)	**1970 - 1995**	684 papers in 4 SSCI journals
Reeves [a (2002)	**1997 - 2001**	597 papers in 3 SSCI journals
Jones Wilson [b (2011)	**2000 - 2009**	2,166 papers in 4 SSCI journals

[a Table footnote a goes here
[b Table footnote b goes here

8. Acknowledgements :

Acknowledgements should appear in a separate section before the reference list.

9. Citations :

Citations in text should follow the author-date method (authors' surname followed by publication year).

- Several studies found... (Barakat et al., 1995; Garfield, 1955; Meho & Yang, 2007).
- In a recent study (Smith & Jones, 2011)...
- Smith and Jones (2011) investigated...

10. Reference List :

Reference list, formatted in accordance with the American Psychological Association (APA) style, should be alpha-betized by the first authors last name.

Journal article

- Author, A., Author, B. & Author, C. (Year). Article title. *Journal Title*, volume(issue), start page-end page.
- Smith, K., Jones, L. J., & Brown, M. (2012). Effect of Asian citation databases on the impact factor. *Journal of Information Science Practice and Theory*, 1(2), 21-34.

Book

- Author, A., & Author, B. (Year). *Book title*. Publisher Location: Publisher Name.
- Smith, K., Jones, L. J., & Brown, M. (2012). *Citation patterns of Asian scholars*. London: Sage.

Book chapter

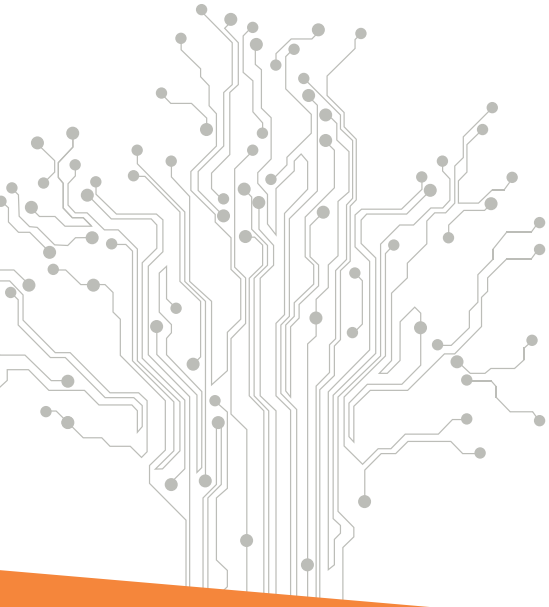
- Author, A., & Author, B. (Year). Chapter title. In A. Editor, B. Editor, & C. Editor (Eds.), *Book title* (pp. xx-xx). Publisher Location: Publisher Name.
- Smith, K. & Brown, M. (2012). Author impact factor by weighted citation counts. In G. Martin (Ed.), *Bibliometric approach to quality assessment* (pp. 101-121). New York: Springer.

Conference paper

- Author, A., & Author, B. (Year). Article title. In A. Editor & B. Editor (Eds.), *Conference title* (pp. xx-xx). Publisher Location: Publisher Name.
- Smith, K. & Brown, M. (2012). Digital curation of scientific data. In G. Martin & L. J. Jones (Eds.), *Proceedings of the 12th International Conference on Digital Curation* (pp. 41-53). New York: Springer.

Online document

- Author, A., & Author, B. (Year). Article title. Retrieved *month day*, *year* from URL.
- Smith, K. & Brown, M. (2010). The future of digital library in Asia. *Digital Libraries*, 7,111-119. Retrieved *May 5, 2010*, from http://www.diglib.org/publist.htm.



JISTaP

Journal of Information Science
Theory and Practice

http://www.jistap.org

KISTI
www.kisti.re.kr

66, Hoegi-ro, Dongdaemun-gu, Seoul, Republic of Korea (ZIP code: 02456)
Tel. +82-2-3299-6102 Fax. +82-2-3299-6067 http://www.jistap.org