How important is software to library and information science research? A content analysis of full-text publications

Xuelian Pan[1], Erjia Yan[2], Ming Cui[1], Weina Hua[1†]

[1]School of Information Management, Nanjing University, Nanjing, China

[2]College of Computing and Informatics, Drexel University, Philadelphia, U.S.A.

Abstract

We investigate the contributions of scientific software to library and information science (LIS) research using a sample of 572 English language articles published in 13 journals in 2008, 2011, 2014, and 2017. In particular, we examine the use and citation of software freely available for academic use in the LIS literature; we also explore the extent to which researchers follow software citation instructions provided by software developers. Twenty-seven percent of the LIS journal articles in our sample explicitly mention and use software. Yet although LIS researchers are becoming increasingly reliant on software that is freely available for academic use, many still fail to include formal citations of such software in their publications. We also find that a substantial proportion of researchers, when documenting software use, do not cite the software in the manner recommended by its developers.

1. Introduction

In the current scientific reward system, scientists' impact is largely assessed via their publication history. This tendency has driven scientists to pursue publications as an end product of their research (Fanelli, 2010; Jacob & Lefgren, 2011; Wang, Liu, Ding, & Wang, 2012). Non-publication outputs, such as data and software, have long been underestimated in comparison with publications (Hafer & Kirkpatrick, 2009; Belter, 2014; Poisot, 2015). However, recent years have witnessed the production of more and

† Corresponding author: huawn@nju.edu.cn, School of Information Management, Nanjing University, NO. 163 Xianlin Road, Nanjing, Jiangsu 210023, China

more non-publication outputs (e.g., scientific data and software), which have played an increasingly important role in advancing scientific theory and practice (Chao, 2011; Belter, 2014; Howison, Deelman, McLennan, Da Silva, & Herbsleb, 2015). As the importance of non-publication outputs is increasingly recognized, some funding agencies, such as the U.S. National Science Foundation and the Higher Education Funding Council for England, have begun to include software, research datasets, and other non-traditional outputs in their consideration of investigators' intellectual contributions (National Science Foundation, 2013; Research Excellence Framework, 2013).

Among the non-publication research outputs, scientific data has attracted the most academic attention because of the widespread recognition that "science is becoming data-intensive and collaborative" (National Science Foundation, 2010; Zacharias, 2010; Tenopir et al., 2011). Many researchers have invested considerable effort into the study of scientific data from numerous perspectives, such as data sharing and reuse, data curation, and data citation (Altman, Borgman, Crosas, & Matone, 2015; Mooney & Newton, 2012; Nelson, 2009; Piwowar & Vision, 2013; Wallis, Rolando, & Borgman, 2013; Witt, Carlson, Brandt, & Cragin, 2009). Compared with scientific data, scientific software has garnered less attention from the academic community and has not been widely valued as an academic contribution. Software has long been considered as supporting service (Howison & Herbsleb, 2014) due to the wide use of commercial software. However, the open source movement has produced vast quantities of free software, much of which has found extensive use in the scientific community in recent years (Huang et al., 2013; Pan, Yan, Wang, & Hua, 2015). Moreover, a substantial proportion of scientists spend a considerable amount of their own research time developing software tools to facilitate their research (Poisot, 2015; Prabhu et al., 2011); in many cases, these tools are then made publicly available (Hannay et al., 2009; Nguyen-Hoan, Flint, & Sankaranarayana, 2010). There is evidence to suggest that these developers are concerned with the use and impact of their software (Trainer, Chaihirunkarn, Kalyanasundaram, & Herbsleb, 2015), and that scientific end users are also interested to know what software others have used (Howison et al., 2015; Huang et al., 2013). Thus, some scholars have begun to investigate the use and impact of software

in scientific publications (e.g., Li, Yan, & Feng, 2017; Pan, Yan, & Hua, 2016).

Some such studies have focused on biology research, where researchers have established the important role that software plays in biological research (Howison & Bullard, 2016; Yang, Rousseau, Wang, & Huang, 2018). Other studies have explored the use and impact of particular software tools, such as R, CiteSpace, HistCite and VOSviewer, in scientific publications; these software tools have likewise been found to have a substantial impact on scientific research (Li, Yan, & Feng, 2017; Pan, Yan, Cui, & Hua, 2018). To date, however, few studies have quantified the impact of scientific software on library and information science (LIS) research. One previous study investigated the proportion of LIS articles containing computing terms in the title, abstract or keywords based on a terminology list, finding that about two thirds of articles post-2000 made mention of computing technologies (Thelwall & Maflahi, 2015). However, this study did not analyze software as a single object, distinct from other technological terms and resources. The present study fills this gap by examining the extent to which scientific software is explicitly mentioned and used in full-text LIS articles.

Citation count, often used to assess the impact of publications and data (Belter, 2014; Cartes-Velásquez & Manterola Delgado, 2014), seems to be suitable for measuring the impact of software as well. However, our previous study has found that more than 40% of software tools used in *PLOS ONE* articles received no formal citations (Pan, Yan, Wang, & Hua, 2015). Howison and Bullard (2016) have likewise found that 56% of software mentions in the biology literature did not include a formal citation. Software "uncitedness" has also been shown to be prevalent in bioinformatics papers (Yang et al., 2018). Taken together, these earlier studies demonstrate that a considerable proportion of software tools are not formally cited in scientific publications. As yet, however, little is known about the extent to which software freely available for academic use is cited in the scientific literature. Earlier evidence has suggested that extrinsic benefits, such as citations and career advancement, motivate scientists to develop and share software (Howison & Herbsleb, 2011; Roberts, Hann, & Slaughter, 2006). A study focusing on software that is freely available for academic use, will thus illuminate the extent to which

developers receive credit for software development and sharing.

Considering that many researchers cite publications but fail to cite software, some scholars have proposed alternative metrics, in addition to citation count, as a means of evaluating the impact of software. They suggest that the number of mentions, downloads, users, registered users, user messages, and user reviews can be used as indicators for measuring this impact (Howison et al., 2015; Pan, Yan, & Hua, 2016; Thelwall & Kousha, 2016; Zhao & Wei, 2017). These indicators are no doubt useful, but accurate data concerning some of them is difficult to collect. For instance, if a software tool, which can be downloaded without payment or registration, is distributed via multiple websites, the user count is hard to obtain. Moreover, some of these indicators may provide a biased picture of the academic impact of scientific software. For example, some users may download a software tool multiple times without ever using it in their research. Faced with such circumstances, other scholars hold that a greater effort must be made to improve the practice of software citation—e.g., by creating software citation principles and developing tools to support software citation (Smith, Katz, & Niemeyer, 2016; Soito & Hwang, 2016). Certainly, much work remains to be done to improve the practice of software citation and the efficacy of research evaluation.

In this study, we extend existing studies on the impact of scientific software to the field of LIS, focusing specifically on the use and citation of software that is freely available for academic use. We aim to answer the following questions:

1. *How important is software to LIS research?*
2. *How is software—in particular, software freely available for academic use—used and cited in LIS research?*
3. *To what extent do LIS researchers cite software as recommended by its developers?*

The answers to the above questions will provide a fuller understanding of the importance of software to scientific research and reveal a more complete and detailed landscape of software citation practices. As the first empirical study focusing on the use of software in the LIS literature, this study will also give a better understanding of the influence of scientific software on LIS specifically. Additionally, this study explores the

discrepancy between LIS researchers' actual citation practices and those proposed as best practices by software developers. Reasons for this lack of consistency are identified, with a view to improving the efficacy of software use and scholarly communications.

2. Methods

2.1. Data Source

Thirteen LIS journals (Appendix A1) were selected from a list of 16 journals used by a previous study on the cognitive structure of LIS (Milojević, Sugimoto, Yan, & Ding, 2011). The set of 16 had itself been highly selective, drawn from a list of important LIS journals rated by American Library Association-accredited education program deans and Association of Research Libraries member library directors (Nisonger & Davis, 2005). Three journals were discarded from the previous list of 16: *Annual Review of Information Science and Technology* was excluded because it ceased publication in 2011; *Reference & User Services Quarterly* and *Library Resources & Technical Services*, were omitted because they publish a large number of practice-oriented papers.

All 3950 research articles published in the 13 journals in 2008, 2011, 2014 and 2017 were downloaded. Book reviews, perspectives, editorials, letters, comments, conference summaries, and other non-research articles were excluded because such articles rarely contain software entities. Appendix A1 shows titles, abbreviations, and article counts for each journal. Our goal was to sample at least 15% of all research papers from each journal in each of the four sampling years. The typical number of research papers per journal per year is between 60 and 90. Thus, we randomly selected 11 articles published in 2008, 2011, 2014, and 2017 from each journal, respectively, as the sample for this study. This left us with a final dataset of 572 full-text articles.

2.2. Content Analysis

Content analysis was employed to examine the use and citation of scientific software in the sample articles. Two coding schemes, shown in Tables 1 and 2, were created based on the work of Howison and Bullard (2016). Our first coding scheme focused on the position and usage of the software, the information about software provided by article

author(s), and the citation of software (Table 1). The second coding scheme focused on whether the software was findable, where it was freely available for academic use, whether software developers provided citation guidelines on the software website, and whether the website included citable works describing the software (Table 2). During the coding process, the coders first identified all software entities actually used in the articles, then coded the identified articles and software entities according to the developed coding schemes and web search results. It should be noted that this study focused on software entities explicitly used rather than merely mentioned in the full text of scientific articles. To see how this distinction was applied, consider the following sentence: "Although science mapping software tools such as CiteSpace and Sci^2 Tool also can deal with bibliometric data downloaded from Web of Science, this study used software VOSviewer to analyse the terms in the title of all selected papers." Here, VOSviewer was coded as software used in the study, whereas CiteSpace and Sci^2 Tool were coded as software mentioned but not used. Software used but not explicitly mentioned in the articles was ignored because it is not feasible to annotate such software correctly. For instance, if a study stated that "a program was written to process the text of each video's title and description," that program was not included in our analysis. A randomly selected sample of 30 articles was coded to assess inter-coder reliability between the two coders, with Cohen's kappa statistics adopted as the measure of reliability. Kappa coefficients for each category were calculated using ReCal2 (http://dfreelon.org/utils/recalfront/recal2/; Freelon, 2010) and found to range from 0.87 to 1, suggesting good agreement (Altman, 1990).

In this article, we counted the number of articles using software, the number of software mentions, and the number of software citations to assess the impact of software on LIS research. The article was adopted as the counting unit in all three instances. For example, one article used VOSviewer to analyze the bibliometric data and mentioned VOSviewer three times in the body text, plus once in the reference list. In this case, the article adds 1 to the count of "articles using VOSviewer", 1 to the count of "VOSviewer mentions", and 1 to the count of "VOSviewer citations".

Table 1. Coding scheme for software mentions and citations.

Code	Description
ArticleID	ID of a particular article that mentions the software. Each article was manually assigned a unique ID before the coders began to annotate the articles. Examples of the ID format include 2008JASIST010, 2011JD041, and 2014JIS002.
Software name	The name of the software, e.g., CiteSpace, Weka, LIBSVM.
Used	Indicates whether the software is used in the research. For instance, the statement that "other software packages (such as CiteSpace or VOSviewer) can also be used to analyze the data" was coded as a mention, but not a use, of both CiteSpace and VOSviewer. (This is, of course, confirmed by reading the article to ensure that use of the two programs is not reported further on.)
Version number	Particular version of the software. For instance, in "SPSS 20.0" and "XLStat 2010", "20.0" and "2010" are version numbers.
URL	Web address of the software. For example, in the sentence "Weka 3.0 (http://weka.wikispaces.com/) was used to analyze the statistical data of each article", "http://weka.wikispaces.com/" is the URL of Weka.
Citation	Indicates whether this paper provides a formal citation of the software in the reference list.
Reference entry	Denotes an entry linked to the software in a reference list.
Reference publication	Denotes citation of a particular publication.
Reference manual	Denotes citation of a specific user guide or manual which is unpublished.
Reference software	Denotes direct citation of a link to the software's website or project name.
Match recommended citation	Denotes whether the authors cite the software as the developers recommended. When the software developers had listed their preferred citation, we compared it with the citation entry.

Table 2. Coding scheme for software attributes.

Code	Description
Software name	The name of the software.
Findable	Indicates whether more detailed information about the software can be found on the Internet, such as the website and online user's guide of the software.
Free	Indicates whether the software is freely available for academic use.
Refer to citation	Denotes whether the software website includes information about how to cite the software.
Provide citable works	Denotes whether there are citable works describing the software (e.g., papers, books, and manuals) on the software website.

3. Results and Discussion

3.1. How important is software for LIS research?

Among the 572 LIS journal articles we surveyed, 153 (27%) explicitly mentioned and used software. Compared to the reported proportion of articles mentioning software (65%) in a previous study on 90 biology papers (Howison & Bullard, 2016), the proportion of articles using software in the field of LIS is small. It should be noted that articles mentioning but not actually using software were not taken into account in our study; this might be one reason for the smaller proportion. By year, the use rate was 27%, 21%, 29%, and 31% (38, 30, 41, and 44 out of 143 articles) for 2008, 2011, 2014, and 2017, respectively. In contrast to the overall proportion of articles using software (range from 5.53% in 2007 to 20.8% in 2016) found in our previous study on papers published in nine Chinese LIS journals (Cui, Pan, & Hua, 2018), the proportion of English LIS journal articles using software is high. Overall, software appears to be more important for research published in English LIS journals than that published in Chinese LIS journals. Moreover, we found that, although the proportion of English LIS journal articles using software has not consistently increased over time, the proportions in 2014 and 2017 are higher than in 2008 and 2011. It should be noted that a certain proportion of authors used software but did not explicitly mention the name of software; seven such articles appeared among the 143 articles published in 2014. This suggests that, in fact, more than 27% of articles in the field of LIS make use of software. We also found that six of the

143 articles published in 2014 used programs developed by the authors; sometimes, LIS researchers must develop software for their studies rather than simply using software developed by others.

Table 3 lists the number of articles using software in each journal. There are marked inter-journal differences: more than 40% of articles in *Library & Information Science Research* (LISR) and *Journal of Academic Librarianship* (JAL) used software, whereas fewer than 20% of articles published in *The Information Society* (IS), *Library Quarterly* (LQ), and *Library Trends* (LT) used software. It is interesting to note that *College & Research Libraries* (CRL) and *Journal of Academic Librarianship* (JAL), two of the four most library-science-oriented journals according to a previous study on the cognitive domains of LIS journals (Milojević, Sugimoto, Yan, & Ding, 2011), have a higher proportion of articles using software than all other journals except *Library & Information Science Research* (LISR) and *Online Information Review* (OIR). The other two journals in this top-four grouping, *Library Quarterly* (LQ) and *Library Trends* (LT), have a smaller proportion of articles using software than all other journals except *The Information Society* (IS). In contrast, 20% of articles published in the two most information-science-oriented journals, *Information Processing & Management* (IPM) and *Journal of the Association for Information Science & Technology* (JASIST), explicitly mentioned and used software, less than that of most of the other LIS journals. Overall, there is no significant difference in the proportion of articles using software between the library-science-oriented journal group (including CRL, JAL, LQ, and LT) and the information-science-oriented group (including IMP, JASIST, JD, and JIS).

Table 3. Proportion of articles using software in each LIS journal by year.

Journal	2008	2011	2014	2017	TotalY	Proportion
CRL	3 (0.27)	6 (0.55)	4 (0.36)	3 (0.27)	16	0.36
IPM	2 (0.18)	0 (0.00)	3 (0.27)	4 (0.36)	9	0.20
IR	3 (0.27)	3 (0.27)	6 (0.55)	1 (0.09)	13	0.30
IS	1 (0.09)	1 (0.09)	2 (0.18)	1 (0.09)	5	0.11
JAL	6 (0.55)	4 (0.36)	4 (0.36)	4 (0.36)	18	0.41

JASIST	1 (0.09)	1 (0.09)	4 (0.36)	3 (0.27)	9	0.20
JD	4 (0.36)	2 (0.18)	2 (0.18)	5 (0.45)	13	0.30
JIS	3 (0.27)	3 (0.27)	3 (0.27)	3 (0.27)	12	0.27
LISR	6 (0.55)	3 (0.27)	4 (0.36)	7 (0.64)	20	0.45
LQ	3 (0.27)	2 (0.18)	1 (0.09)	1 (0.09)	7	0.16
LT	1 (0.09)	0 (0.00)	1 (0.09)	0 (0.00)	2	0.05
OIR	3 (0.27)	4 (0.36)	4 (0.36)	5 (0.45)	16	0.36
SCI	2 (0.18)	1 (0.09)	3 (0.27)	7 (0.64)	13	0.30
TotalJ	38	30	41	44	153	0.27
P	0.27	0.21	0.29	0.31	0.27	/

Note. The number in parentheses is the ratio of the number of articles using software to 11; TotalY indicates the total number of articles using software published in the four sampling years for each journal; Proportion = 153/(11×4×13) if TotalY equals 153, else Proportion = TotalY/(11×4); TotalJ indicates the total number of articles using software published in the 13 journals in each year; P = 153/(11×4×13) if TotalJ equals 153, else P = TotalJ/(11×13).

A total of 75 distinct software entities, each of which is explicitly mentioned and used, are identified from the 572 articles and they are mentioned 218 times. Among these 75 entities, the statistical software SPSS is the most frequently used, with 52 mentions—suggesting that about 9% of LIS articles use SPSS. Other statistical software packages such as SAS, STATA, Minitab, and XLStat were also used in these articles. Data storage and processing tools, such as Excel, Access, and SQL Server, are also frequently used in LIS research. In addition to these general-purpose tools, bibliometric mapping software (e.g., Bibexcel, BICOMS, CiteSpace, Sci2, Thomson Data Analyzer, VantagePoint and VOSviewer), social network analysis packages (e.g., Netdraw, NodeXL, Pajek and Ucinet), structural equation modeling tools (e.g., AMOS, LISREL and SmartPLS), qualitative data analysis packages (e.g., ATLAS.ti and Nvivo) and data mining/natural language processing tools (e.g., LIBSVM, MALLET, NLPIR and Weka) have all been adopted by LIS researchers. Seven distinct bibliometric mapping tools are represented in our sample, more than for any other identifiable type of software tool.

3.2. How is software used and cited in LIS research?

Of the 75 pieces of software identified above, 69 could be found on the Internet; 6 could not. After manually checking all 69 pieces of software, we found that 33 were commercial (the *commercial group*) and 36 were freely available for academic use (the *freeware group*). Commercial tools were used a total of 156 times, whereas the freeware tools were used a total of 56 times. This result reveals that, on average, commercial software tools are more frequently used in LIS research, even though they are fewer in number than freeware tools. Our finding is different from that of a related study (Huang et al., 2013), in which commercial software tools were much less frequently used by bioinformatics researchers than what we here term freeware tools. Moreover, as shown in Table 4, instances of freeware use in in LIS research have increased from 6 in 2008 to 27 in 2017. At the same time, we find that the proportion of mentions of freeware has increased from 0.12 in 2008 to 0.48 in 2017, suggesting that such software is becoming more and more important to LIS research.

Table 4. Summary of software mentions by year.

Year	2008	2011	2014	2017	Total
Software free for academic use	6	8	15	27	56
All types of software	52	39	54	73	218
Proportion	0.12	0.21	0.28	0.48	0.26

Note. Proportion indicates the rate of freeware mentions and is calculated as (Mentions of freeware/Mentions of all types of software).

Location and version information are useful for readers who wish to find a given software package; however, only 6% and 23%, respectively, of the 218 software mentions include website and version information in their text. Seventy-two percent of the mentions provide only the name of the software, with no further information given. This suggests that descriptive information such as website and version number were frequently overlooked when LIS researchers documented their software use. This finding

is in accordance with our previous study on the use, citation, and diffusion of bibliometric mapping software tools.

Furthermore, only 18% of the 218 software mentions include references. The citation rates for sampling years 2008, 2011, 2014, and 2017 are 0.08, 0.15, 0.15, and 0.29, showing an overall increasing trend. This may reflect the extensive efforts made to improve software citation practices during the past few years, such as the development of citation standards and the creation of tools to support software citation.

The cited references related to the software are examined in greater depth, with the result shown in Table 5. Among the 39 references to software, 25 (64%) cite a related publication, 14 (36%) directly cite software (with 10 including a link to the software's website), and none cite a user manual. Thus, LIS researchers seem most likely to cite a related publication when making a citation to software they have used. Our finding contrasts with the observation of Yang et al. (2018) that biologists preferred to cite software directly. Table 5 also shows that researchers who published in the sole informetrics journal in our sample (SCI) were most likely to make a formal citation to software, while authors who published in the most library-science-oriented journals (CRL, JAL, LQ, and LT) did not include formal software citations.

Table 5. Software citation rate for articles published in each LIS journal.

Journal	Mentions	Citations	Citation rate	References to software		
				Reference publication	Reference software	Reference manual
CRL	18	0	0.00	0	0	0
IPM	14	3	0.21	3	0	0
IR	15	1	0.07	0	1	0
IS	5	2	0.40	0	2	0
JAL	22	0	0.00	0	0	0
JASIST	11	1	0.09	0	1	0
JD	25	9	0.36	7	2	0
JIS	21	8	0.38	4	4	0
LISR	28	2	0.07	2	0	0
LQ	12	0	0.00	0	0	0

LT	3	0	0.00	0	0	0
OIR	26	4	0.15	2	2	0
SCI	18	9	0.50	7	2	0
Total	218	39	0.18	25	14	0

3.3. To what extent do LIS researchers cite software as recommended by software developers?

We calculated separate citation rates for the commercial and freeware groups using SPSS (SPSS, version 20; IBM Corp., Armonk, NY), with the result shown in Table 6. We found that 11% (95% CI: 0.06–0.16) of commercial software tools received citations, while 38% (95% CI: 0.24–0.51) of software tools freely available for academic use received citations. There was a statistically significant difference in citation rate between the two groups (two-tailed Pearson's chi-squared test, $p < 0.05$). One possible reason is that freeware developers are more likely to furnish information on how to cite their software and to provide citable publications on their websites. Indeed, just 7 (21% of 33) of the commercial software tools include related publications on their websites, while 29 (81% of 36) of the freeware packages provide such publications.

Table 6. Citation rate for four groups of software tools.

Group	Mentions	Citations	Citation rate	95% confidence interval
Commercial software group	156	17	0.11	0.06-0.16
Freeware group	56	21	0.38	0.24-0.51
Mention-of-citation group	28	13	0.46	0.27-0.66
Nonmention-of-citation group	28	8	0.29	0.11-0.46

We next turned our attention to freeware developers' provision of citation guidelines. The commercial software tools were ignored because their developers are seeking to sell them, rather than to enhance their academic reputation (Howison & Bullard, 2016). Among the 36 freeware tools, 15 (42%) included citation guidelines on their websites. Moreover, some developers provide information on how to cite their software in places

other than the software homepage. For instance, the creators of CiteSpace and Pajek provide citation recommendations in the software interface and the user manual, respectively. These findings provide evidence that a considerable proportion of academic freeware developers are concerned with the citation of their software.

We further classified the freeware group into two subgroups according to whether the developers provide citation information on their website. The first subgroup is the *mention-of-citation group* (15 pieces of software, which were used 28 times); the remaining pieces of software form the *nonmention-of-citation group* (21 pieces of software, also used 28 times) (see Table 6). The mention-of-citation group (citation rate: 0.46) has a higher software citation rate than the nonmention group (0.29). On the one hand, the results suggest that providing software citation instructions has some benefit for improving citation practices; on the other hand, more than 50% of software mentions did not include any formal citations, even though the developers provided citation information on their websites.

We now focus on the mention-of-citation group: the 15 software tools whose websites include citation guidelines. As shown in Table 7, we find that in 8 (53% of 15) cases, the developers recommend citing a related publication; in 6 (40% of 15) cases, direct citation of the software is proposed; and in 1 (7% of 15) case, citation of a user manual is recommended. This reveals a lack of consistency in the form of citation recommended by developers, suggesting a possible reason for the current diversity of software citation practices. We also find that only two software developers recommend mentioning the explicit version of the software, and seven developers suggest mentioning the software's website. Although DOI has been increasingly recommended for software citation and can be easily obtained by submitting the code to a digital repository such as Zenodo (Smith et al., 2016; Soito & Hwang, 2016), we find that none of the developers provide a citable DOI in the software citation instructions. In total, the 15 software tools were used 28 times in 22 papers, but only 10 (67% of 15) tools were cited and only 13 citations made. It should be noted that 8 software citations occurred in 2017, 3 citations in 2014, and the remainder (2) in 2011. Even among the tools that received citations, only 6 were cited as recommended by their developers. In other words, 21% (6 of 28) of

software mentions were accompanied by the recommended citations. This means that nearly 80% of LIS researchers did not follow existing developer guidelines when they documented their use of software.

Table 7. Basic statistics for software tools which are freely available for academic use and have software citation instructions on their website.

Software	Recommended citation	Version	URL	Mentions	Citations	Cite as recommended
AntConc	Cite software	Yes	No	1	0	0
BibExcel	Cite a publication	No	No	2	0	0
BICOMS	Cite a publication	No	No	1	1	1
GeoDa	Cite a publication	No	No	1	0	0
LIBSVM	Cite a publication	No	Yes	2	2	1
MALLET	Cite software	No	Yes	1	1	1
NetDraw	Cite a publication	No	No	2	1	0
NodeXL	Cite software	No	Yes	4	2	0
plyr	Cite software	No	Yes	1	0	0
Publish or Perish	Cite software	No	Yes	1	1	1
R	Cite a user manual	No	Yes	4	2	0
Sci2	Cite software	No	Yes	2	1	1
Stanford Parser	Cite a publication	Yes	No	1	1	1
Webometric Analyst	Cite a publication	No	No	2	0	0
Weka	Cite a publication	No	No	3	1	0
Total	/	/	/	28	13	6

Note. "Recommended citation" indicates the type of citation target recommended by software developers; "Version" indicates that the developer recommendations included the citation of version number information; "URL" indicates that the recommendations included a link to the software's website; "Cite as recommended" indicates the number of citations that the users made to the software as recommended.

4. Conclusion

This study examines the importance of software to LIS research as well as the use

and citation of software that is freely available for academic use in the scientific literature. Moreover, this article explores the degree to which software citation instructions are promulgated and followed. We first selected a sample of 572 articles from the 3950 research articles published in 13 LIS journals in 2008, 2011, 2014, and 2017, then performed content analysis to identify software packages as well as characteristics of these software packages.

Results showed that nearly 30% of the LIS research articles explicitly mentioned and used software, with increasingly heavy reliance on software in recent years. Moreover, articles in each of the 13 LIS journals used one or more software tools in the four sampling years, though there were marked inter-journal differences in the extent of such use (the proportions of articles using software range from 0.05 to 0.45). These findings demonstrate the importance of software to LIS research: although it is generally agreed that software plays an important role in scientific research (Pan, Yan, & Hua, 2016), we still know little about the extent to which this is true in the field of LIS. Our results provide evidence that software is instrumental to about one third of LIS research. These findings answer the first of the three research questions posed at the beginning of this study.

Turning to our second question, the results also revealed that LIS researchers mentioned and cited software in diverse ways. Only 6% of researchers provided website information, 23% provided version information, and more than 70% provided no further information than the name of the software in the text. Fewer than 20% of software mentions included a formal citation in the reference list. Even among the LIS researchers who made a formal citation to software used in their research, there is an observable inconsistency in the choice of what to cite: 64% of researchers cite publications related to software, while 36% directly cite the software itself. In this respect, the software citation practices of LIS researchers are similar to those of researchers in other fields (e.g., Yang et al., 2018). Another notable finding of the current study was that the software citation rate shows an overall increasing trend. This may be related to the ongoing effort to standardize, facilitate, and improve software citation practices. In addition, the increasing proportion of freeware tools used in LIS research may be driving the rise in citation rate,

because software that is freely for academic use was itself more likely to receive citations.

Considering that the developers of commercial software are more interested in earning monetary rewards than in making scientific contributions (Howison & Bullard, 2016), we further examined the use and citation of software freely available for academic use, whose developers are more interested in making a contribution to science and building an academic reputation. The results showed that LIS researchers are becoming increasingly reliant on this latter class of software, which we referred to above as freeware or academic freeware. However, more than 60% of the mentions of such software did not include a formal citation in the reference list. Moreover, although more than 40% of the developers of freeware tools provided software citation instructions on their websites, they held different views on what to cite: among the developers who made a recommendation, 53% of suggested users cite publications related to the software, 40% proposed that users directly cite the software, and 7% recommended users cite a user manual. The diversity of these proposed software citation instructions might be a reason for the inconsistency of software citation practices.

We also found that the average citation rate of the freeware tools with official citation instructions was higher than for tools without such instructions, though not to a statistically significant extent. In a previous study on the citation of the Protein Data Bank, the authors found that users tend to cite the data repository as proposed by the citation instructions (Huang, Rose, & Hsu, 2015). This contrasts with our results: even freeware tools with official citation instructions were, more often than not, mentioned without a formal citation. Moreover, a considerable proportion of researchers did not follow the software citation instructions proposed by the developers even if they did make a formal citation of some kind.

This study has a few limitations. In particular, it focuses on 13 journals and thus cannot accurately represent software use and citation behaviors in the field of LIS overall. These journals were selected from a ranked journal list based on LIS expert opinion, and all were indexed in Web of Science. Our previous study, however, has provided evidence that Chinese LIS journals are less likely to rely on software (Cui, Pan, & Hua, 2018) than

those studied here. In addition, the small sample size may influence some of the results; further studies with larger sample sizes are needed to confirm these preliminary findings. The small number of mentions of the freeware tools with official citation instruction is another limitation to be kept in mind when interpreting the findings presented here.

Despite the above limitations, this study explores three important aspects of software use and citation in LIS research: the overall importance of software to the field, the use and citation of software that is free for academic use, and the extent to which LIS researchers cite software as suggested. Our findings furnish a fuller understanding of the importance of software to scientific research and shed light on a significant lack of consistency—both in the citation instructions provided by software developers and in the software citation practices of LIS researchers. Many interesting questions, however, remain unanswered. Future research will examine how software users choose what to cite, why users do not cite software as recommended by its developers, and how academic freeware developers arrive at such recommendations to begin with.

Acknowledgments

This work was funded by the National Natural Science Foundation of China (Grant No. 71704077).

Reference

Altman, D. G. (1990). *Practical statistics for medical research*. Boca Raton, FL: CRC Press.

Altman, M., Borgman, C., Crosas, M., & Matone, M. (2015). An introduction to the joint principles for data citation. *Bulletin of the American Society for Information Science and Technology*, *41*(3), 43–45.

Belter, C. W. (2014). Measuring the value of research data: A citation analysis of oceanographic data sets. *PLOS ONE*, *9*(3).

Cartes-Velásquez, R., & Manterola Delgado, C. (2014). Bibliometric analysis of articles published in ISI dental journals, 2007–2011. *Scientometrics*, *98*(3), 2223–2233.

Chao, T. C. (2011). Disciplinary reach: investigating the impact of dataset reuse in the earth sciences. *Proceedings of the ASIST*, *48*.

Cui, M., Pan, X., & Hua, W. (2018). Software usage and citation in the field of library and information science in China. *Journal of Library Science in China*, *44*(235), 66–78.

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PLOS ONE*, *5*(4).

Freelon, D. (2010). ReCal: intercoder reliability calculation as a web service. *International Journal of Internet Science*, *5*(1), 20–33.

Hafer, L., & Kirkpatrick, A. E. (2009). Assessing open source software as a scholarly contribution. *Communications of the ACM*, *52*, 126.

Hannay, J. E., MacLeod, C., Singer, J., Langtangen, H. P., Pfahl, D., & Wilson, G. (2009). How do scientists develop and use scientific software? In *Proceedings of the 2009 ICSE workshop on software engineering for computational science and engineering, SECSE 2009* (pp. 1–8). New York, NY: ACM.

Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., … Nishioka, T. (2010). MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry*, *45*(7), 703–714.

Howison, J., & Bullard, J. (2016). Software in the scientific literature: problems with seeing, finding, and using software mentioned in the biology literature. *Journal of the Association for Information Science and Technology*, *67*(9), 2137–2155.

Howison, J., & Herbsleb, J. D. (2011). Scientific software production: incentives and collaboration. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 513–522). New York, NY: ACM.

Howison, J., & Herbsleb, J. (2014). *The sustainability of scientific software: ecosystem context and science policy*. Working Paper. University of Texas at Austin. Retrieved from http://james.howison.name/pubs/HowisonHerbsleb-Sustainability.pdf

Howison, J., Deelman, E., McLennan, M. J., Da Silva, R. F., & Herbsleb, J. D. (2015). Understanding the scientific software ecosystem and its impact: current and future measures. *Research Evaluation*, *24*(4), 454–470.

Huang, X., Ding, X., Lee, C. P., Lu, T., Gu, N., & Hall, S. (2013). Meanings and boundaries of scientific software sharing. In *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW)* (pp. 423–434). New York, NY: ACM.

Huang, Y. H., Rose, P. W., & Hsu, C. N. (2015). Citing a data repository: a case study of the protein data bank. *PLOS ONE*, *10*(8), e0136631.

Jacob, B. A., & Lefgren, L. (2011). The impact of research grant funding on scientific productivity. *Journal of Public Economics*, *95*(9–10), 1168–1177.

Li, K., Yan, E., & Feng, Y. (2017). How is R cited in research outputs? Structure, impacts, and citation standard. *Journal of Informetrics*, *11*(4), 989–1002.

Milojević, S., Sugimoto, C. R., Yan, E., & Ding, Y. (2011). The cognitive structure of Library and Information Science: analysis of article title words. *Journal of the American Society for Information Science and Technology*, *62*(10), 1933–1953.

Mooney, H., & Newton, M. (2012). The anatomy of a data citation: discovery, reuse, and credit. *Journal of Librarianship and Scholarly Communication*, *1*(1), eP1035.

Nelson, B. (2009). Data sharing: empty archives. *Nature*, *461*(7261), 160–163.

Nguyen-Hoan, L., Flint, S., & Sankaranarayana, R. (2010). A survey of scientific software development. *Proceedings of the 2010 ACM-IEEE international symposium on Empirical software dngineering and measurement - ESEM '10*, 1.

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. *College Research Libraries*, *66*, 341–377.

National Science Foundation. (2010). Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans. Retrieved from https://www.nsf.gov/news/news_summ.jsp?cntn_id=116928

National Science Foundation. (2013). *GPG summary of changes*. Retrieved from https://www.nsf.gov/pubs/policydocs/ pappguide/nsf13001/gpg_sigchanges.jsp

Pan, X., Yan, E., Cui, M., & Hua, W. (2018). Examining the usage, citation, and diffusion patterns of bibliometric mapping software: a comparative study of three tools. *Journal of Informetrics*, *12*(2), 481–493.

Pan, X., Yan, E., & Hua, W. (2016). Disciplinary differences of software use and impact in scientific literature. *Scientometrics*, 1–18.

Pan, X., Yan, E., Wang, Q., & Hua, W. (2015). Assessing the impact of software on science: a bootstrapped learning of software entities in full-text papers. *Journal of Informetrics*, *9*(4), 860–871.

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175.

Poisot, T. (2015). Best publishing practices to improve user confidence in scientific software. *Ideas in Ecology and Evolution*, *8*, 50–54.

Prabhu, P., Kim, H., Oh, T., Jablin, T. B., Johnson, N. P., Zoufaly, M., ... Ghosh, S. (2011). A survey of the practice of computational science. In *SC'11: Proceedings of 2011 international conference for High performance computing, networking, storage and analysis* (pp. 1–12). New York, NY: IEEE.

Research Excellence Framework. (2013). *Output information requirements*. Retrieved from http://www.ref.ac.uk/about/guidance/submittingresearchoutputs/

Roberts, J. A., Hann, I. H., & Slaughter, S. A. (2006). Understanding the motivations, participation, and performance of open source software developers: a longitudinal study of the Apache projects. *Management Science*, *52*(7), 984–999.

Smith, A. M., Katz, D. S., & Niemeyer, K. E. (2016). Software citation principles. *PeerJ Computer Science*, *2*, e86.

Soito, L., & Hwang, L. J. (2016). Citations for software: providing identification, access and recognition for research software. *International Journal of Digital Curation*, *11*(2), 48–63.

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., … Frame, M. (2011). Data sharing by scientists: practices and perceptions. *PLOS ONE*, *6*(6), 1–21.

Thelwall, M., & Kousha, K. (2016). Academic software downloads from Google Code. *Information Research*, *21*(1). Retrieved from http://files.eric.ed.gov/fulltext/EJ1094576.pdf

Thelwall, M., & Maflahi, N. (2015). How important is computing technology for library and information science research? *Library & Information Science Research*, *37*(1),

42–50.

Trainer, E. H., Chaihirunkarn, C., Kalyanasundaram, A., & Herbsleb, J. D. (2015). From personal tool to community resource: What's the extra work and who will do it? In *Proceedings of the 18th ACM conference on Computer supported cooperative work & social computing* (pp. 417–430). New York, NY: ACM.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. *PLOS ONE*, *8*(7), e67332.

Wang, X., Liu, D., Ding, K., & Wang, X. (2012). Science funding and research output: A study on 10 countries. *Scientometrics*, *91*(2), 591–599.

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing data curation profiles. *International Journal of Digital Curation*, *4*(3), 93–103.

Yang, B., Rousseau, R., Wang, X., & Huang, S. (2018). How important is scientific software in bioinformatics research? A comparative study between international and Chinese research communities. *Journal of the Association for Information Science and Technology*. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.24031

Zhao, R., & Wei, M. (2017). Impact evaluation of open source software: an Altmetrics perspective. *Scientometrics*, *110*(2), 1017–1033.

Appendix A1: Summary of thirteen LIS journals used as data sources.

Journal title	Abbreviation	2008	2011	2014	2017	Total
College & Research Libraries	CRL	30	30	41	49	150
Information Processing & Management	IPM	113	64	52	77	306
Information Research	IR	56	46	71	128	301
Information Society	IS	26	23	26	26	101
Journal of Academic Librarianship	JAL	57	60	77	65	259
Journal of Documentation	JD	43	43	53	72	211
Journal of Information Science	JIS	50	52	63	52	217
Journal of the Association for Information Science & Technology	JASIST	185	186	184	202	757
Library & Information Science Research	LISR	30	37	25	34	126
Library Quarterly	LQ	18	16	28	28	90
Library Trends	LT	35	41	33	27	136
Online Information Review	OIR	52	48	52	64	216
Scientometrics	SCI	128	218	346	388	1,080

Note. "Total" indicates the total number of research articles published in 2008, 2011, 2014, and 2017. The title of the *Journal of the Association for Information Science & Technology* was changed in 2014 from *Journal of the American Society for Information Science AND Technology*.