Qualitative and Quantitative Methods in Libraries (QQML) 5: 365- 377, 2016

Beyond the Download: Issues in Developing a Secondary Usage Calculator

Carol Tenopir[1], Lisa Christian[2], Robert Anderson[3], Lorraine Estelle[4], Suzie Allard[5], and Dave Nicholas[6]

1University of Tennessee, Knoxville
2University of Tennessee, Knoxville
3University of Tennessee, Knoxville
4Project COUNTER
5University of Tennessee, Knoxville
6CIBER

Abstract: Since 2002, Project COUNTER has led the way in developing and maintaining systems of measurement for download counts. While these counts have often been used as a proxy measure in determining journal and article value for libraries and publishers, they miss an important post-download secondary usage factor – namely, that of sharing. Likewise, altmetrics, while accounting for the impact of social media, misses some aspects of sharing as distribution often occurs via email. This creates difficulty in quantifying an exact measure of use. One aim of the Beyond Downloads project was to develop a calculator for measuring total digital usage – including sharing. Through an examination of a range of sharing systems, we identified the most commonly used platforms for sharing scholarly articles, while an international survey provided data on access, download, saving, and sharing behavior. Survey results indicated that a range of sharing patterns can be estimated, but post-download usage often is too skewed to establish exact calculations. Therefore, in lieu of a sharing calculator, we derived ranges of sharing patterns at a confidence level of 95%. These ranges vary dependent on many factors. Furthermore, our project highlighted the primary issues in developing a secondary usage calculator –namely, the lack of global standards in sharing data and the necessity to extend a survey data approach into a longitudinal study. As a consequence, we recommend a two-fold approach going forwards: 1) data-based approach in which a COUNTER-like universal standard for sharing data is developed and adopted across multiple platforms; 2) survey-based approach in which a longitudinal study is administered to a multi-disciplinary online community.

Keywords: scholarly communication, usage studies, secondary usage, user behaviour, calculator

Received: 15.4.2016 Accepted: 21.5.2016 ISSN 2241-1925

1. Introduction

Traditionally, publishers and libraries have relied on usage statistics to monitor the downloading of scholarly articles from academic journals and to compare download volume by platform and title. However, these statistics do not account for post-download usage of articles. Secondary usage derived from the sharing of downloaded articles is widespread, occurring through formal methods, those products and services specifically designed for the sharing of scholarly content, and through informal methods, whose primary function is not the archiving or dissemination of scholarly publications. Calculating this secondary usage could provide critical data toward a better understanding of overall article value and impact (Tenopir, Hughes, Christian, Allard, & Nicholas, 2015).

Research into literature on the sharing of scholarly work suggests that there is a gap in understanding the extent of full-text article sharing and a lack of effective means to quantify this sharing. Previous studies have identified the attitudes and practices of article sharing by scholars, as well as the stages of sharing (Cheng, Ho, & Lau, 2009; Brown, 2010; Acord & Harley, 2013; Fitzpatrick, 2012). Additionally, many studies examined the various methods in which scholars share their work (Acord & Harley, 2013; Fitzpatrick, 2012). Altmetrics, that is methods for measuring impact beyond the traditional citation count, help to bridge an understanding between the usage data and the actual, real-world influence of scholarly work (Lapinski, Piwowar, & Priem, 2013; Roemer & Borchardt, 2012). However, calculation of post-download article sharing remains a relatively unexplored area.

One of the Beyond Download project's objectives was to develop practical ways to estimate total digital article usage from downloads and non-download usage. In order to meet these objectives, the study aimed to develop a method of calculating or estimating secondary usage. Analysis focused on key questions related to counting secondary usage. Applying a Confidence Interval to survey findings provides an estimated "range of sharing," with a lower bound and upper at a 95% level of confidence. Combined with download numbers, this range could estimate an approximate level of actual, post-download usage.

2. Methodology

In the process of developing our survey, we identified formal and informal methods of sharing scholarly articles. The search focused on platforms most likely used by scholars for research and teaching. Formal methods of sharing are platforms developing with sharing in mind; they include learning management systems such as Blackboard, reference management systems such as Endnote, file sharing and storage or cloud services, and research social networks such as Mendeley and ResearchGate. Informal methods are those systems not specifically developed for article sharing, but used for it

nonetheless. Informal methods include: Twitter, blogs, email, Facebook, LinkedIn, and other general social networks.

Elsevier distributed an email invitation to authors who have contributed to any of their journals. This invitation included a link to an online questionnaire. The mailing list totaled 32,956 authors and we received 1000 responses to at least one question for a response rate of 3.03%. The anonymous survey opened 31 October 2014 and closed 16 January 2015. Respondents were allowed to leave the survey at any time, skip questions, or were timed out automatically if they began the questionnaire and did not complete it. The study was approved by the University of Tennessee Institutional Review Board for Human Subjects.

Respondents were asked 34 questions regarding their download, saving, and sharing behaviors, as well as personal demographics. The demographic questions allowed us to see how those issues may vary according to age, subject/discipline, and country of origin, highest degree held, rank/position, and experience with research group work. We used a "critical incident technique" to ask respondents to recall their most recent incidence of sharing. This allowed respondents to focus on more details of their sharing behavior rather than attempting to recall more general sharing behaviors (Flanagan, 1954).

In order to gain a more detailed picture of the usage of different platforms for sharing, we asked respondents, "Thinking back to the last scholarly article that you published, please estimate how many times and/or with how many people you shared FULL-TEXT articles" for each platform listed (Table 1). However, for this article, we focus only "how many times" articles are shared in the process of writing an article.3

Table 1. Sharing of Full-Text Articles

	Times Shared		**95% CI Range**	
	M	**SD**	**Lower**	**Upper**
Email	10.50	29.89	7.98	13.02
Internal Networks	3.92	15.02	2.26	5.58
Cloud Service	3.13	11.00	1.94	4.32
Reference Management Software	1.93	10.64	0.68	3.18
Learning Management Software	2.64	15.64	0.82	4.47
Research social networks	10.84	107.49	-0.34	22.01
General social networks	2.41	14.16	0.78	4.04
Other	12.20	1400.06	2.75	27.15

Applying a Confidence Interval, derived from Sample Size, Mean, and Standard Deviation, we find a narrow average of post-download usage in some areas. For example, with a 95% confidence, we can estimate that respondents in the sciences are sharing articles via email between 5.73 and 10.6 times. Unfortunately, this level of confidence cannot be expressed with every subject or means of sharing due to the small sizes of certain variables.1

In order to work around these limited sample sizes, we re-coded certain demographic characteristics. We collapsed subject disciplines into broader subject areas, and grouped ages by decade. We extracted lower and upper ranges of sharing from statistical analysis derived from these re-coded demographics.2

3. Limitations

Only 1.6% of respondents are from humanities/fine arts due to Elsevier's mailing list, which is comprised mostly of researchers in the sciences, medical sciences, social sciences, and engineering / computer sciences / mathematics (E/CS/M).

In addition, only 2.4% of the respondents are under age 30, so any results from this range must be considered carefully before making any definitive statements on the sharing behaviors of respondents less than 30 years of age.

4. Findings

4.1 Informal Methods of Sharing

Email

Examining sharing through email (Table. 2), we find that for the last scholarly article that respondents published, they shared articles roughly ten times by email (M=10.5), with a range of sharing falling roughly between eight to 13 times. Sharing based on the subject discipline of the respondent, is consistent with that of the overall average.

Comparisons via age show a wider variation of results. Email shares in the Under 30 category (M=3.33) fell well below the survey average. Respondents in their 30s also share less via email than the average. Those 60 and Over share more than the average.

Table 2. Number of Articles Shared Via Email *

	Mean	Lower	Upper
Email (Overall)	10.50	7.98	13.02
Subject Discipline			
Sciences	8.17	5.73	10.6

Medical Sciences	12.79	6.78	18.8
Engineering/Computer Sciences/Mathematics	13.43	4.19	22.67
Social Sciences	9.3	6.54	12.05
Age Group			
Under 30	3.33	0.84	5.82
Age 30-39	10.44	2.73	18.15
Age 40-49	9.14	5.22	13.07
Age 50-59	10.46	6.92	13.99
60 and Over	14.4	8.65	20.15

*Mean and 95% Confidence Interval

Internal Networks

Overall respondents indicate an average of four shares per last article published with a range between two and six (Table 3). Estimated sharing via internal networks by scientists is slightly less than the overall estimate with a relatively comparable range. In the medical sciences, E/CS/M and social sciences categories, sharing through internal networks is compatible with that of the average and the sharing behaviors of scientists.

Internal networks shows some variation within age groups. Respondents under 30 years share significantly less through internal networks than the overall average. Respondents at least fifty years old share more through internal networks than any other age group.

Table 3. Number of Articles Shared Via Internal Network *

	Mean	**Lower**	**Upper**
Internal Network (Overall)	3.92	2.26	5.58
Subject Discipline			
Sciences	2.99	0.11	5.87
Medical Sciences	3.55	0.89	6.21
Engineering/Computer Science/Mathematics	5.99	1.06	10.91
Social Sciences	3.85	0.61	7.09
Age Group			
Under 30	0.33	-0.44	1.1
Age 30-39	3.14	-0.13	6.41

Age 40-49	2.35	0.23	4.47
Age 50-59	5.36	1.14	9.58
60 and Over	5.3	0.8	9.81

*Mean and 95% Confidence Interval

General Social Networks

We found some differences between disciplines in sharing articles through general social media (Table 4). For instance, engineers/computer scientists/mathematicians and social scientists share more through general social media. In fact, their averages are higher than the overall sharing average for general social media.

In the age categories, there is some inconsistency. No respondent under 30 years reported sharing via general social networks. Respondents in their 30s and 40s share less than the overall average, with respondents in their 50s more than twice the average. Sharing in the 60 and over group is less than half of the overall average.

Table 4. Number of Articles Shared Via General Social Networks *

	Mean	**Lower**	**Upper**
General Social Networks (Overall)	2.41	0.78	4.04
Subject Discipline			
Sciences	1.34	0.01	2.68
Medical Sciences	1.01	0.21	1.82
Engineering/Computer Sciences/Mathematics	3.41	-0.24	7.06
Social Sciences	3.85	-2.21	9.91
Age Group			
Under 30	0	0	0
Age 30-39	1.85	-0.47	4.17
Age 40-49	1.96	0.65	3.27
Age 50-59	5.95	-1.56	6.06
60 and Over	0.82	-0.14	1.79

*Mean and 95% Confidence Interval

Other

Among the various sharing methods, we also included a catch-all "other" category for all other sharing methods not listed in the survey. When choosing this method, we asked respondents to identify these other methods. They included sharing printed copies by hand or mail, by direct transfer of files onto computers, written or printed references shared by hand, word-of-mouth, and through USB drives or other mobile devices. Sharing through other informal methods is mostly consistent with that of the identified informal methods, with slight variations in ranges of sharing by subject and age category (Table 5). Full-text articles are shared 0.85 through 2.09 times by other means not listed in the survey. In the sciences and medical sciences, the number of times shared via other methods is slightly higher than that of the overall average. In the age categories, respondents under 30 years did not report sharing through other means. Those 60 and Over (M=2.1) share significantly less.

Table 5. Number of Articles Shared Via Other Informal Methods *

	Mean	Lower	Upper
Other (Overall)	12.20	2.75	27.15
Subject Discipline			
Sciences	21.88	-18.26	62.02
Medical Sciences	1.56	0.01	3.1
Engineering/Computer Sciences/Mathematics	22.93	-20.81	66.67
Social Sciences	1.23	0.09	2.37
Age Group			
Under 30	0	0	0
Age 30-39	19.52	-17.6	56.64
Age 40-49	20.75	-17.81	59.31
Age 50-59	1.26	0.28	2.24
60 and Over	2.1	0.06	4.15

*Mean and 95% Confidence Interval

4.2 Formal Methods of Sharing

Cloud

When sharing articles via cloud services, respondents from the sciences, medical sciences, and social sciences show very similar results to the overall survey results (Table 6). Those in the E/CS/M category share significantly more via cloud services. As with some informal sharing methods, respondents under

30 years remain largely under-represented. Otherwise, sharing is similar to the average.

Table 6. Number of Articles Shared Via Cloud Services*

	Mean	Lower	Upper
Cloud (Overall)	3.13	1.94	4.32
Subject Discipline			
Sciences	1.79	0.64	2.95
Medical Sciences	3.29	0.89	5.68
Engineering/Computer Sciences/Mathematics	6.61	1.61	11.61
Social Sciences	3.35	0.23	6.47
Age Group			
30 and Under	0.2	-0.1	0.5
Age 30-39	3.79	0.69	6.9
Age 40-49	3.52	1.32	5.73
Age 50-59	2.47	0.48	4.45
60 and Over	3.1	1.18	5.02

*Mean and 95% Confidence Interval

Reference Management Software

Respondents in the sciences and social sciences report sharing via reference mangers slightly less than the average, while sharing by respondents in the medical sciences and engineering/computer science/mathematics report sharing twice as often as the overall average (Table 7). Respondents in their 40s and 50s share higher than the average, with the latter sharing almost twice the average.

As with subject category results, breaking down sharing via reference management software into age show some fluctuation. Respondents in their 30s and 60 and Over are slightly below the overall average. Those 60 and Over share significantly less than average.

Table 7. Number of Articles Shared Via Reference Management Software*

	Mean	Lower	Upper
Reference Manger (Overall)	1.93	0.68	3.18

Subject Discipline			
Sciences	0.65	0.1	1.2
Medical Sciences	2.5	-0.34	5.34
Engineering/Computer Sciences/Mathematics	2.84	-0.52	6.19
Social Sciences	2.22	-1.19	5.63
Age Group			
Under 30	0	0	0
Age 30-39	1.71	-0.64	4.05
Age 40-49	2.19	-0.13	4.51
Age 50-59	3.5	-0.53	7.53
60 and Over	0.59	-0.04	1.21

*Mean and 95% Confidence Interval

Learning Management Software

Respondents in the sciences report lower amounts of sharing through learning management software than the overall average (Table 8). Conversely, respondents in the medical and social sciences report slightly higher than average amounts of sharing. Respondents in the E/CS/M and medical sciences share twice as much as the overall Mean. Those in their 30s report significantly more sharing than the overall results. Results indicate that respondents in their 40s and 50s share less.

Table 8. Number of Articles Shared Via Learning Management Software*

	Mean	**Lower**	**Upper**
Learning Manger (Overall)	2.64	0.82	4.47
Subject Discipline			
Sciences	0.25	0	0.51
Medical Sciences	1.17	0.04	2.3
Engineering/Computer Sciences/Mathematics	4.58	-0.52	6.19
Social Sciences	5.31	-0.97	11.59
Age Group			
Under 30	0	0	0
Age 30-39	4.76	-0.67	10.19
Age 40-49	1.07	0.16	1.98

Age 50-59	1.18	0.36	2
60 and Over	2.62	-1.59	6.82

*Mean and 95% Confidence Interval

Research Social Networks

Respondents in the sciences, medical sciences and E/CS/M report sharing less through research social networks than the overall average (Table 9). However, respondents in the social sciences (M=28.25) category report significantly greater amounts of sharing. In the age categories, respondents in their 50s share the most through research social networks (M=5.17), but their average is still less than the overall average (M=10.84).

Table 9. Number of Articles Shared Via Research Social Networks*

	Mean	Lower	Upper
Research Social Networks (Overall)	10.84	-0.34	22.01
Subject Discipline			
Sciences	3.66	1.91	5.41
Medical Sciences	4.35	2.44	6.25
Engineering/Computer Sciences/Mathematics	9.74	0.64	18.83
Social Sciences	28.25	0.51	9.03
Age Group			
Under 30	0.3	-0.05	0.72
Age 30-39	4.74	1.69	7.8
Age 40-49	4.73	-22.88	89.41
Age 50-59	5.17	1.64	8.71
60 and Over	3.94	1.77	6.11

*Mean and 95% Confidence Interval

5. Discussion and Conclusion

Several factors contribute to the scarcity of reliable data on post-download sharing. A lack of global standards in regards to sharing data makes quantifying sharing problematic; hence, a data-based approach to quantifying the extent of post-download sharing of articles may not be viable. The great bulk of measurable activity for a typical article takes place on the original publisher's site and that it is a significant challenge for publishers to obtain reliable, consistent data from even the most important and best managed research

management systems (RMSs), suggesting it is unlikely that this could be done for a wide range of such services on an ongoing basis. It is difficult to obtain data on authors' apparently widespread sharing of articles via email and cloud services. Due to the variation of methods of sharing articles, data obtained is likely to be out of date rather quickly. The lack of global standards for data reporting by RMSs and the enormity of data that would have to be collected, processed and weighed make quantifying sharing a logistical improbability.

After identifying formal and informal methods of sharing we attempted to contact publisher services in order to obtain any relevant data or insights regarding usage and sharing. These publishers included six major journal publishers (ACS Publications, Nature Publishing Group, PLOS, Spring, Taylor & Francis, and Wiley), one aggregator, and HighWire Press. Most publishers expressed an interest in quantifying the extent of sharing. Additionally, they were able to provide helpful commentary on the value and reliability of the data on post-download article activity that they obtain from a range of sources.

We suggest two approaches moving forward:

1) Data-based approach: This should be confined to data that is stable, reliable and can be obtained on a regular basis. It falls into two categories:
    a) First, usage data from publishers, subject repositories and institutional repositories, where there are global standards in place and there are processes being developed for collecting and consolidating such data
    b) Second, citation data, which is comprehensive and readily available.
2) Survey-based approach: using an online community that covers all scholarly disciplines, participants' sharing behavior could be monitored on a regular basis. This method can take into account all types of sharing activity, as well as changing behavior. While the results may not, initially, be reliably quantitative, more precise quantitative data will become available as the community develops and more granular monitoring becomes possible.

6. Acknowledgements

Elsevier B.V. supported this research project. We thank our colleagues from Elsevier who provided insight and expertise that greatly assisted the research. We thank Peter Shepherd, formerly Executive Director of COUNTER, for assistance with identifying formal and informal methods of sharing scholarly materials, and Suzan Ali Saleh, University of Tennessee graduate research assistant, for survey analysis. We would also like to thank CIBER Research Ltd and Project COUNTER.

References

Tenopir, C., Hughes, G., Christian, L., Allard, S., Nicholas, D., Watkinson, A, Woodward, H., Shepherd, P., & Anderson, R. (2014). To Boldly Go Beyond Downloads:

How Are Journal Articles Shared and Used? Charleston Conference Proceedings.doi: 10.5703/1288284315614
Acord, S. K., & Harley, D. (2013). Credit, time, and personality: The human challenges to sharing scholarly work using Web 2.0. New Media & Society, 15(3), 379-397. doi: 10.1177/1461444812465140
Brown, S. (2010). Socialized Scholarship: It Starts with Us. English Studies in Canada, 36(4), 9-12.
Cheng, M.-Y., Ho, J. S.-Y., & Lau, P. M. (2009). Knowledge Sharing in Academic Institutions: a Study of Multimedia University Malaysia. Electronic Journal of Knowledge Management, 7(3), 313-324.
Fitzpatrick, K. (2012). Giving It Away: Sharing and the Future of Scholarly Communication. Journal of Scholarly Publishing, 43(4), 347-362.
Flanagan, J. C. (1954). The critical incident technique. Psychological bulletin, 51(4), 327-358. doi: 90/10.1037/h0061470
Lapinski, S., Piwowar, H., & Priem, J. (2013). Riding the crest of the altmetrics wave How librarians can help prepare faculty for the next generation of research impact metrics. College & Research Libraries News,74(6), 292-300.
Roemer, R. C., & Borchardt, R. (2012). From bibliometrics to altmetrics A changing scholarly landscape. College & Research Libraries News, 73(10), 596-600.
Suri, V., Sourabh, P., Nihar, T., & Netti, K. (2013). Spatial outlier detection using improved z-score test. International Journal of Engineering Science and Technology, 5(12), 1962.
Rousseeuw, P. J., & Leroy, A. M. (2005). Robust regression and outlier detection (Vol. 589). John Wiley & Sons.
Yang, S., & Berdine, G. (2016). Outliers. The Southwest Respiratory and Critical Care Chronicles, 4(13), 52-56.

---

1 It should be noted that in some instances the lower range of the confidence interval falls below 0. It can be assumed that the range of sharing in these cases should start at 0, as one cannot share a negative number of times.

2 These same demographic categories were assessed for divergence using z-score analysis in order to reduce skewing and kurtosis from extreme outliers. The z-score serves to measure the divergence of results by indicating the number of standard deviations a particular element is from the mean. This measure provides a mechanism to determine the magnitude by which an observation deviates from the remainder of the dataset, and if found to be large enough, the irregular element can be deemed an outlier (Suri, Sourabh, Nihar, & Netti, 2013; Rousseeuw & Leroy, 2005). We identified outliers as those variables falling above an absolute value of 3 and truncated them from the overall survey results. We took precautions to avoid bias by adjusting the cutoff criteria to the parameters of each cell (Rousseeuw & Leroy, 2005). Then, we re-ran the truncated data and compared the results from adjusted batches to the unaltered results.

3 The findings are presented unaltered as outliers are expected with the given sample size. Outliers cannot be attributed to the misreading of instruments or measurements. Additionally, the research in this area is unprecedented and normal distribution cannot be assumed. The outliers might be "valid extreme observations due to random variability" (Yang & Berdine, 2016) as such they represent the randomization of the survey data. We

had initial concerns that the outliers would skew the results, but once compared to truncated findings, any level of skewing was deemed too minimal to justify outlier removal.