Multilingual Scholarship: Non-English Sources and Reference Management Software

Adam H. Lisbon

PII: S0099-1333(17)30118-0
DOI: https://doi.org/10.1016/j.acalib.2017.12.001
Reference: ACALIB 1876

To appear in: *The Journal of Academic Librarianship*

Received date: 20 March 2017
Revised date: 22 November 2017
Accepted date: 13 December 2017

Please cite this article as: Adam H. Lisbon , Multilingual Scholarship: Non-English Sources and Reference Management Software. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Acalib(2017), https://doi.org/10.1016/j.acalib.2017.12.001

Title:

Multilingual Scholarship: Non-English Sources and Reference Management Software

Author:

Adam H. Lisbon
Japanese and Korean Studies Librarian
Assistant Professor

Affiliation:

University of Colorado Boulder
University Libraries
184 UCB
1720 Pleasant Street
University of Colorado
Boulder, CO 80309-0184

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Multilingual Reference Management

Abstract

Introduction

Research happens across borders and languages, but software for managing and citing information is not designed to consider the linguistic diversity of the world's resources. Reference management software (RMS) such as Endnote, Mendeley, and Zotero, among many others help keep track of information sources for research projects. The strength of these programs is based on automatically pulling metadata from online sources to neatly show the title, author, date of publication, etc. of books, articles , and more. RMS should then automatically output metadata into proper citation and bibliographic formats, saving researchers time. However, when sources are not in English, and especially when they are not written in Roman characters (i.e. the Latin alphabet), RMS have no way to faithfully store the original vernacular language side by side with transliterations/translations.

For example, the author 村上春樹 must be written as "Murakami Haruki" for the sake of English speakers who do not read Japanese. Likewise, his book 海辺のカフカ would need a transliteration of "Umibe no Kafuka," and a translation of "Kafka on the Shore." Publishers in the English speaking world often expect some combination of vernacular, transliteration, or translation when non-English sources are cited. Current RMS do not allow for representation of a work across multiple languages or scripts.

This article investigates how multilingual researchers (MLR) use or avoid RMS. Through an in-depth survey, this study intends to understand MLRs perception of RMS and what changes such software will have to make in their functionality in order to accurately represent non-English information sources.

Literature Review

Research on RMS breaks down into several major categories: technical comparisons; LIS professionals and their knowledge of RMS, educating users; user attitudes on RMS and how they use it; and most recently, whether RMS even fits into the workflow of someone's research process.

Multilingual issues are rarely mentioned in RMS research. In 2011 and 2013, Francese investigated awareness and usage at the Universities of Tallinn, Estonia and Torino, Italy. Wu and Chen (2012) asked a small focus group of students at National Taipei University about their experiences with RMS. Sarrafzadeh and Hazeri (2014) investigated use and awareness at Persian Gulf University, Iran. Ram and Paul Anbu K. (2014) conducted a similar study in India. Throughout all of these studies, concerns over the representation of materials not in the native language(s) were never noticed, considered, or asked. Only Melles and Unsworth (2015) quoted one participant in their study that

> "...I work, particularly for my primary documents, essentially entirely in French, and EndNote automatically capitalises and French titles don't capitalise. And I ... tried EndNote and found it so incredibly frustrating because I would have to go back and manually change everything. I gave up on it."

Noguchi (2009) compared RefWorks, EndNote, and Zotero's ability to import Japanese scripts from four major databases: First Search, the Library of Congress Catalog, and four Japan specific databases: WebCat, the National Diet Library (of Japan) Catalog, CiNii, and Magazine Plus. At the time of the study,

only RefWorks could automatically pull Japanese language data from FirstSearch. The CiNii database was also able to interact with these RMS with varying degrees of automation. For all other instances, manual entry was necessary. However, this wasn't a true multilingual study as it only explored citation management in one language, Japanese.

While Library and Information Science literature has not addressed multilingual research and RMS, the TeX/LaTeX/BibTeX community has been exploring solutions to multilingual typesetting and bibliographies. Harders (2002) and Hufflen (2009) each explored these implementations of BibTex. However, these documents are technical, with a steep learning curve. Developing a knowledge base of TeX, LaTex, and BibTex is not practical for most RMS users. Mead & Berryman (2010), Francese (2011, 2013), Hicks & Sinkinson (2015), and Melles & Unsworth (2015) all point to learning curves and/or time as a major reason why users don't develop a strong foundation using RMS. Childress (2011), Ram and Paul Anbu K. (2014) further point out that even LIS professionals and graduate students don't always have a strong grasp of RMS software.

The first true discussion of multilingual RMS did not arise until Bennett (2013) published *Citations Out of the Box*, introducing Juris-M. At the time it was called Multilingual Zotero (MLZ). The book did not explore RMS use, or the needs of MLRs, but was a technical explanation of how Juris-M works. It explained the development of the software. Built entirely off of Zotero's open access code, it was developed to allow parallel metadata fields for vernacular, transliterated, and translated titles, authors, publishers, etc. This allowed users to accurately represent their information sources "as they are" while simultaneously recording Romanized transliterations and/or translations that publishers may require when citing those materials[1].

In studies on RMS use, Juris-M has not been mentioned. In both of Francese's studies, Melles and Unsworth, Sarrafzadeh and Hazeri, Wu and Chen, all showed EndNote having the highest rate of awareness and use. Francese's 2013 study and Melles & Unsworth found that not using any RMS was the second most common experience, 24% and 29% in their respective samples. Melles & Unsworth pointed out that further research on RMS should not be premised on adoption rates, but on the user's needs and how they conduct research.

Adoption of RMS is driven by the recommendation of colleagues (Francese 2013) or professors' recommendations to their students (Wu and Chen 2012). EndNote, having been around since the 1980s, has a user base large enough to make word-of-mouth an effective means of propagating its use. Resistance to experimenting with newer programs like Zotero and Mendeley was pointed out by one of Melles and Unsworth's (2015) participants:

> "I just look at [other software] and think, I've spent all these hundreds of hours and money on EndNote and I'll have to start again and learn Zotero. I've looked at Mendeley, and they all promise the world and then when you go to work them, it's not as easy…"

[1] Juris-M also has special metadata fields for legal citations. Legal citation management is primarily how it is marketed. Bennett is a Law professor at the University of Nagoya, Japan. Hence his particular background on the issue of multilingual legal citation management. The project to develop Juris-M began in 2009, when it was originally known as Multilingual Zotero (MLZ). The name was changed in 2013 to emphasize the software's ability to cite legal documents, and to distance itself from the original Zotero software. Juris-M is an independent project.

Francese (2013) notes that 87% of participants never asked for nor received support on how to use RMS. Even when training is provided, adoption is low. Melles and Unsworth (2015) noted one participant who was unable to grasp the usefulness of RMS even though they attended a training session. Francese (2013) found the two most used features were editing citations and inserting them into papers, a feature often referred to as "write and cite." Melles and Unsworth (2015) further noted RMS users were unaware of even this core feature. Hicks and Sinkinson (2015) found Mendeley users were unaware of its social networking capabilities, one of the particular features Mendeley markets to set itself apart from other RMS.

Childress (2011) addressed the idea of teaching RMS as a set of best practices. Melles and Unsworth (2015) further concluded that RMS must be taught in the context of existing practices, with special recognition that undergraduate students are still developing their own research skills and styles. Hicks & Sinkinson (2015) took this notion further, concluding that RMS are only one part of a larger nexus of digital literacies and digital scholarship. These concepts in turn are creating fundamental shifts in perceptions of what scholarship and the research process are.

Methods

For the purposes of this study, a multilingual researcher (MLR) was defined as anyone who uses information sources that are not in English, but publishes their findings in English. Participants did not need to exclusively use non-English sources, and qualified as conducting "multi-lingual" research even if they had used only a few information sources in foreign languages. Students who produce class papers or dissertations also qualified under this definition. On average, participants conducted research in three languages in addition to English. Because publication requirements in the English speaking world were of primary interest, researching with information sources in English and publishing in another language or any other combination of languages was not considered.

Data on how MLRs use or do not use reference management software (RMS) was gathered through a survey of primarily multiple choice and Likert scale questions, with some open text questions to understand participants' unique situations and perceptions. Open text fields were analyzed for recurring themes and normalized. There were a total of 172 participants. The survey branches at key points to identify several sub-populations including: researchers primarily using Roman script sources (N=82), those using primarily non-Roman script sources (N=90), RMS users (N=61), and non RMS users (N=111).

In this study, "Roman scripts" were defined as those that use the Latin alphabet, and primarily include Romance and Slavic languages. Non-Roman scripts are those that use some other alphabet, such as Russian, Greek, Yiddish, or Hindi, or those that use characters like Chinese and Japanese. There are several languages that cannot neatly fit into these categories. Vietnamese, for example, used a non-Roman script before colonization, but French colonizers replaced the native writing system with one based on the Latin alphabet. Azerbaijani is another example, it was written in Cyrillic during the Soviet era, and today it is written with Latin letters.

The survey was open to anyone who wished to take it, with snowball sampling employed to gather participants across disciplines and languages. Subject specialist librarians at the CU-Boulder libraries shared the survey across the humanities, social sciences, and natural sciences. Participants were encouraged before and after the survey to share it with other colleagues, their departments, email lists, etc. 172 complete responses were recorded that met the definition of an MLR who publishes in English.

Demographics

Participants in this sample represented many academic disciplines. Most were in the humanities and social sciences. Despite trying to recruit researchers in the natural sciences, only two participated. Disciplines with significant representation included art & art history, classics, Jewish/Yiddish studies, linguistics, and history. On average, participants conducted research in three languages other than English. The top 15 most interacted with languages are listed in figure 1.

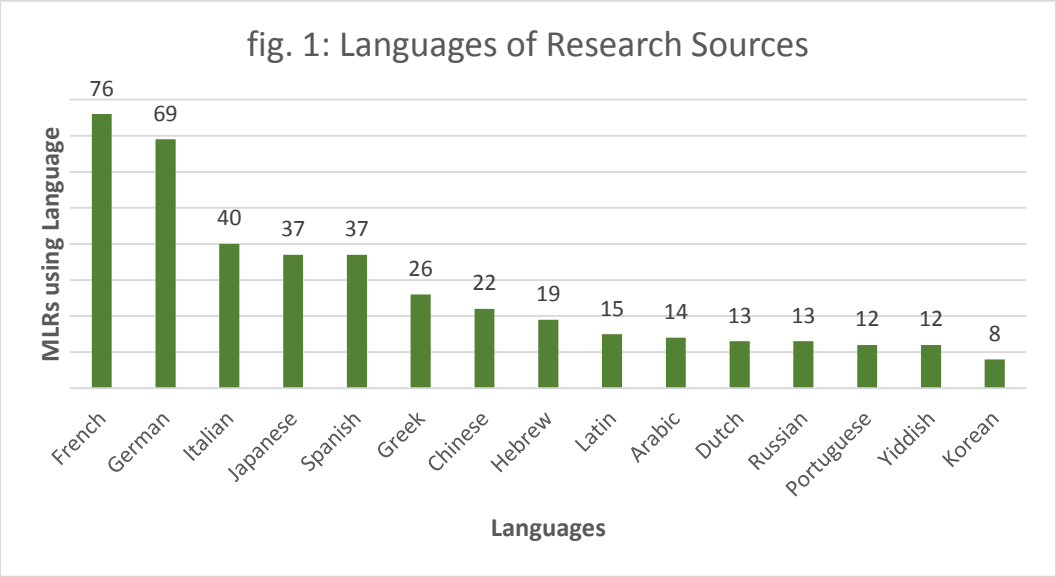
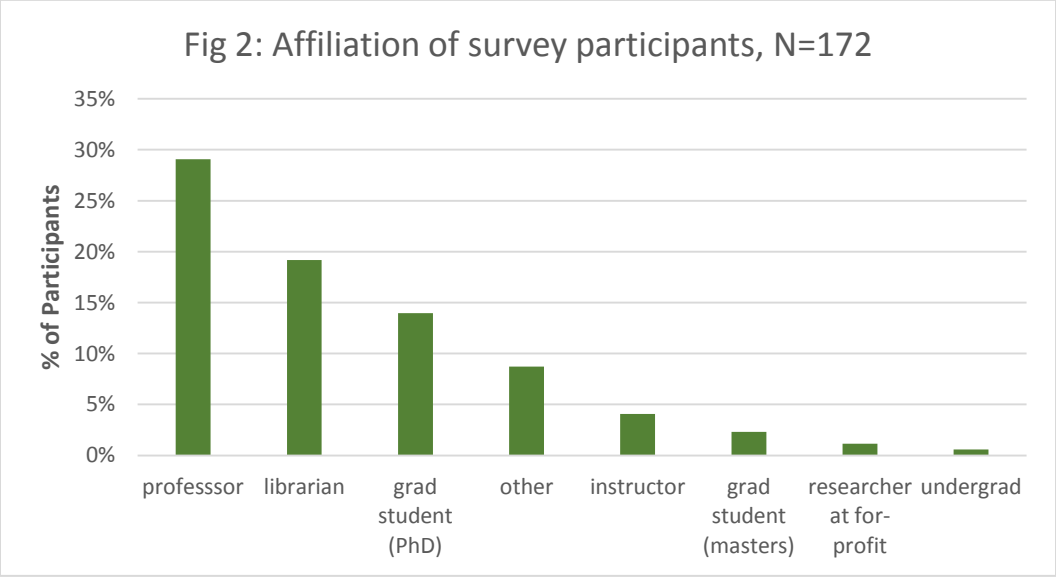
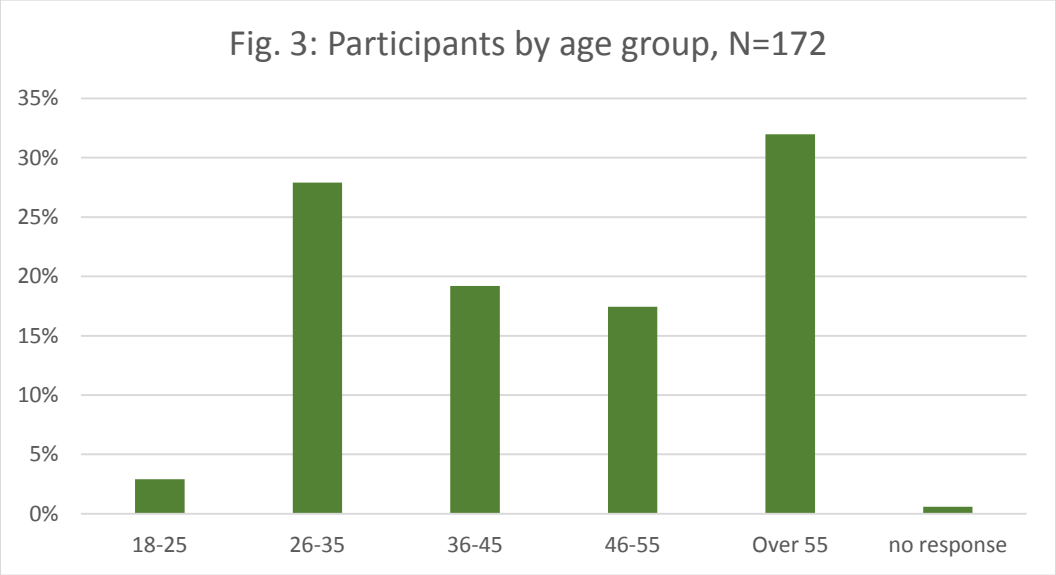


Figure 2 shows how participants identified their professions/roles.



Professors, librarians, and PhD graduate students represented the majority of participants. Consequently, the sample reflected an academic scholarly perspective. Participants were also asked their age (see figure 3), with all but one volunteering this information. The over 55 demographic constituted the largest sub population (n=xyz), followed by the 26-35 bracket. 18-25 year olds had the

lowest response rate. Finally, this sample overwhelmingly represented the experience of people living in the US (n=141).



Multilingual Researchers and Their Citation Needs

In order to understand trends in RMS use, establishing expectations for citing non-English sources was essential. To understand this, participants were asked "When you submit/hand in/publish your research (or write class papers/theses), are you required to translate the titles of your sources?"

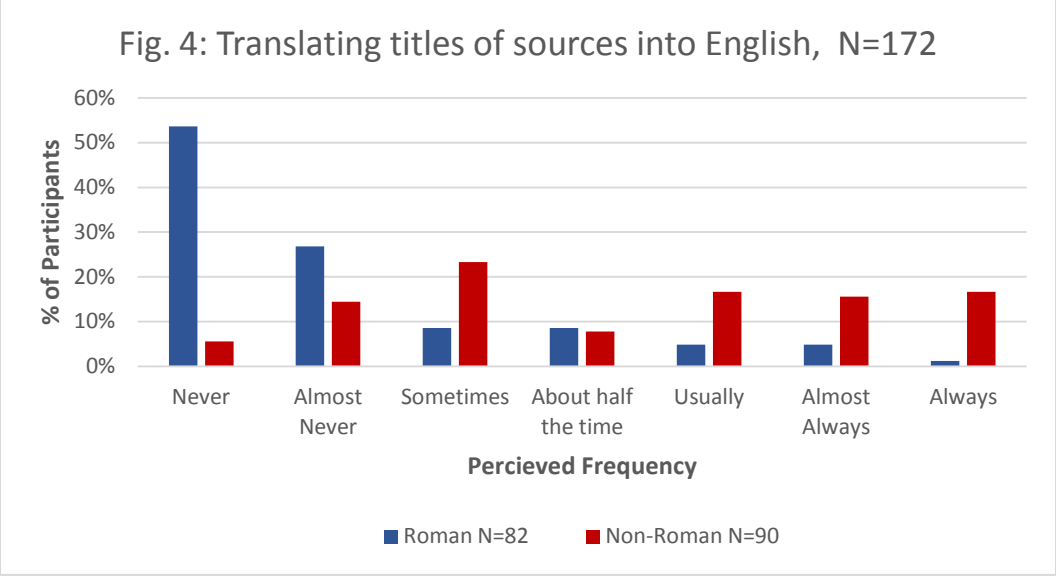
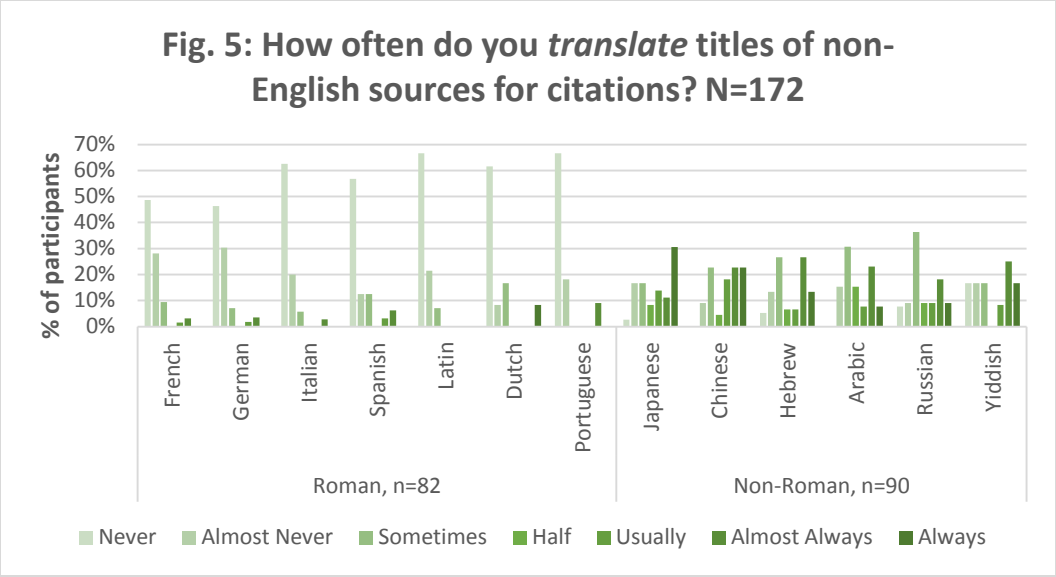
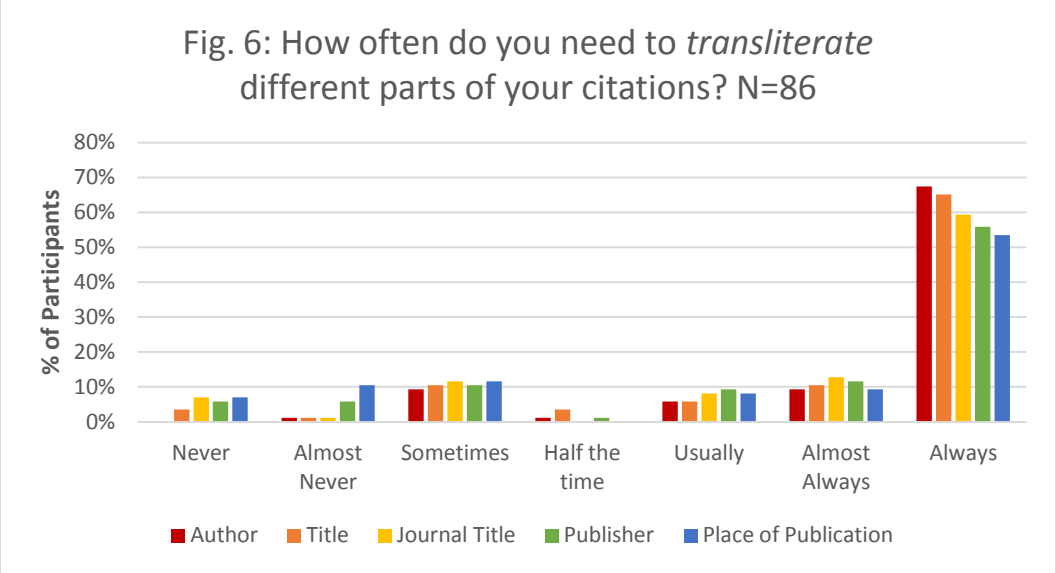


Figure 4 shows Non-Roman MLRs are expected to translate the titles of their sources with greater frequency than their Roman MLR counterparts. Over 50% of Roman MLRs have never done so. Most Non-Roman MLRs translate their sources at some point. All participants answered this question regardless of using RMS or not. This trend is broken down in greater detail by language in figure 5.

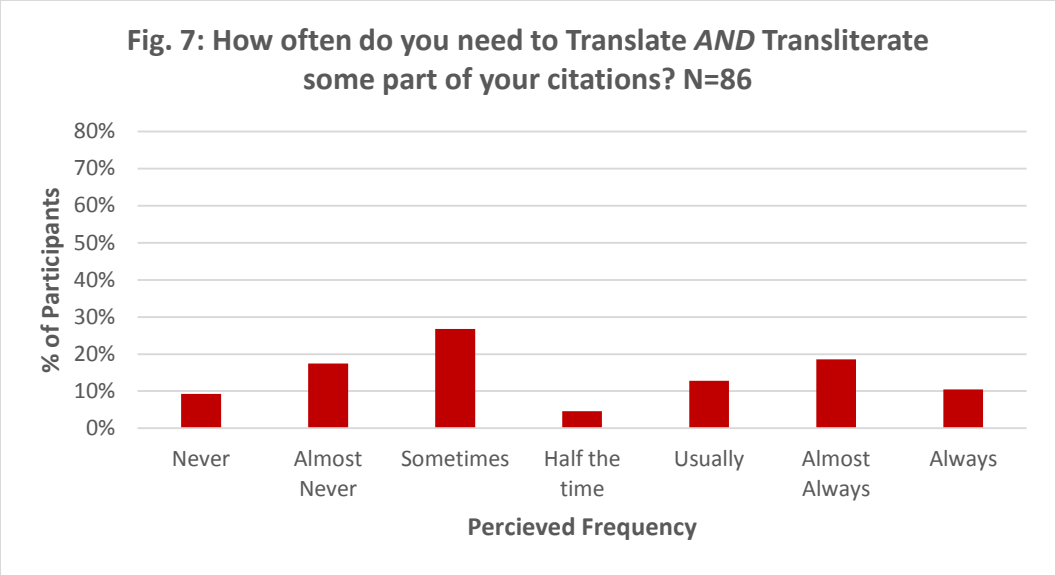


Researchers using Non-roman sources translate titles far more often. Translating still occurs when using Roman script sources, but many working in those languages never have to do so. Sources in ideographic scripts (Chinese and Japanese) are translated more often, while other languages using non-Latin alphabets are translated less.

Transliteration, unlike translation, only applies to MLRs working with non-Roman sources. In addition to titles, citations typically require the author, journal title, publisher, and place of publication to be written out in the Latin alphabet. Figure 6 shows that the vast majority of non-Roman MLRs transliterate their sources.



Though not as common, non-Roman MLRs provide both translations AND transliterations of their citations. In Figure 7, the numbers are more dispersed across frequency, but most have created citations that include translations and transliterations at some point.



When Roman and non-Roman MLRs were asked how important it is to record transliterations and translations in addition to vernacular data on their sources, figure 8a and 8b show non-Roman MLRs placed more value in doing so. This correlates with the data in figures 4, 5, and 6, showing that translations and particularly transliterations are expected by publishers. These perceived values point to the functionality that MLRs need RMS to have.

Fig. 8a: Roman Users Percieved Importance for Storing Vernacular, Transliterated, and Translated Bibliographic Information, n=31

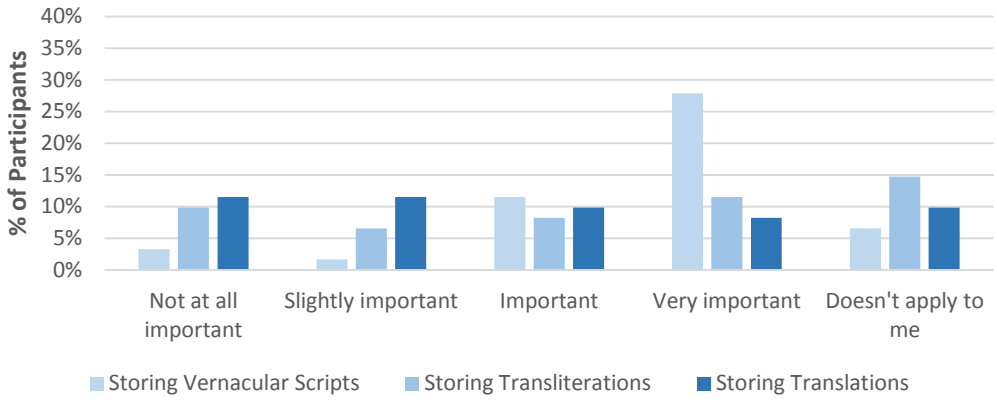
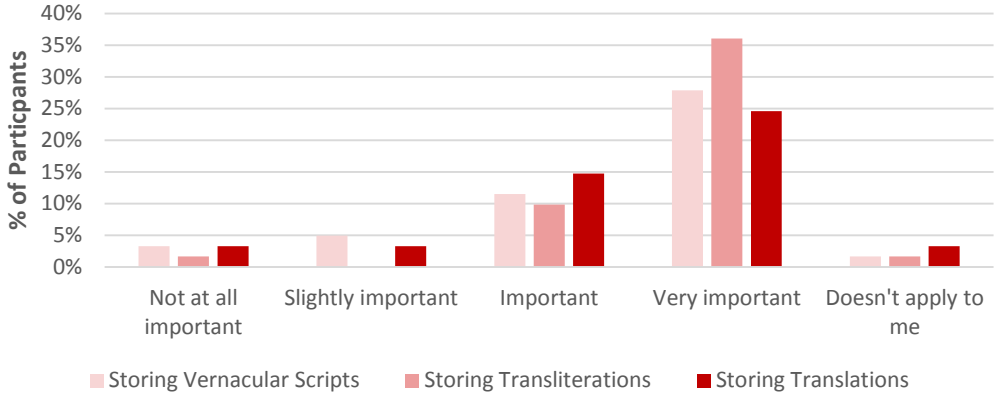
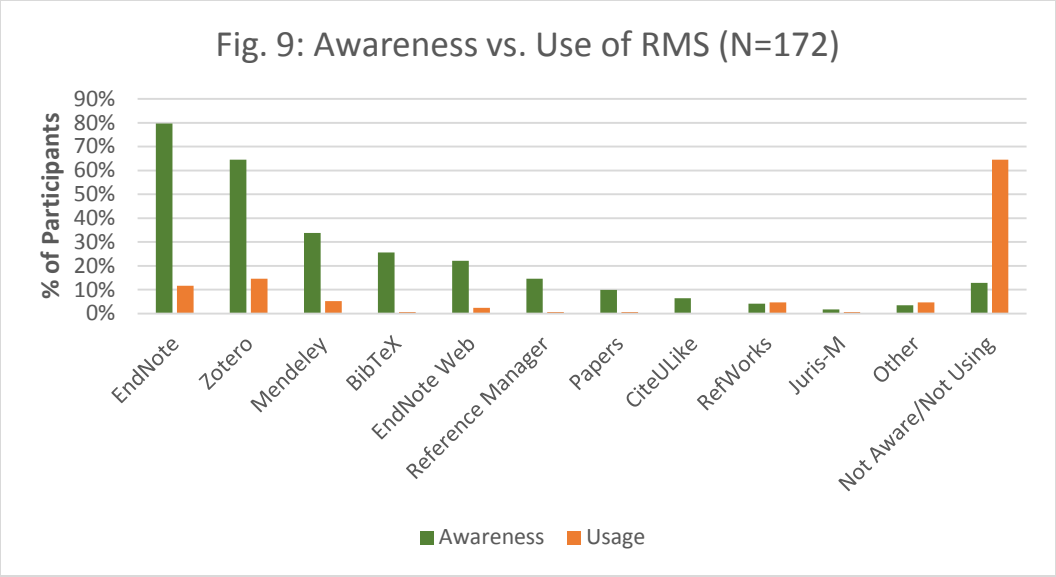


Fig 8b: Non-Roman Users Percieved Importance for Storing Vernacular, Transliterated, and Translated Bibliographic Information, n=30



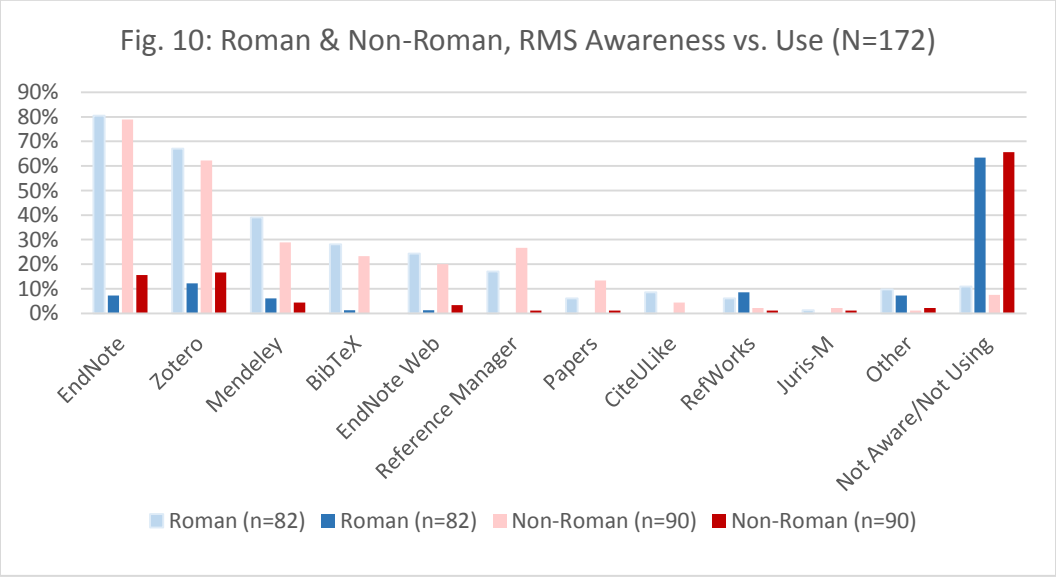
Multilingual Researchers Use and Non-Use of RMS

MLRs need to accurately represent information sources in transliteration, translation, or both. Before examining specific multilingual issues related to RMS, it is essential to assess the current state of RMS use by MLRs. Figure 9 shows the awareness vs. the usage of various RMS.



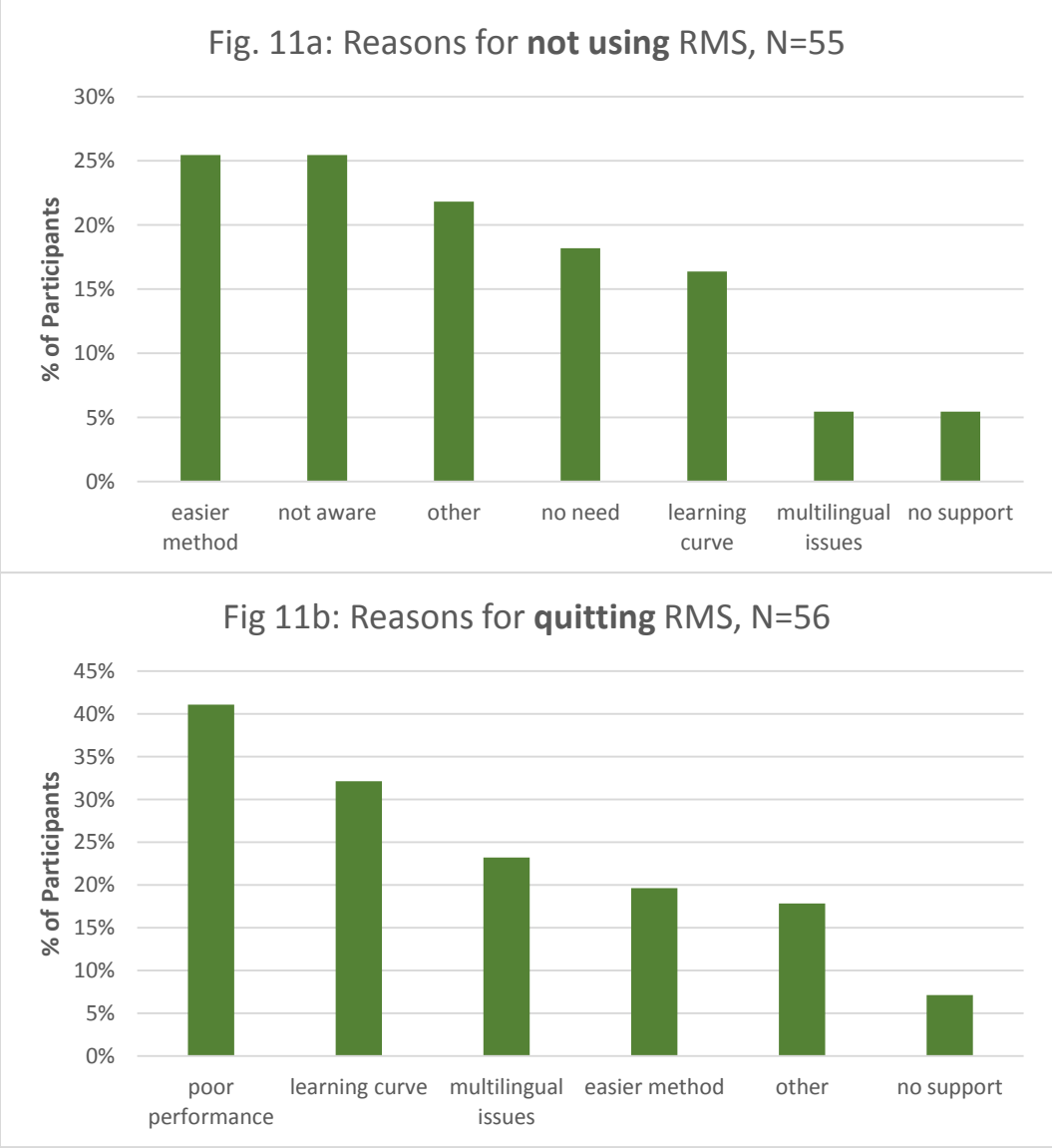
Most participants were aware of more than one RMS, while 13% (22) of participants were unaware of any. 65% (111) did not use any RMS at all. The rate of non-use is high compared to other studies on RMS. Francese (2013), and Melles and Unsworth (2015) measured rates of non-use in the samples at 24% and 29% respectively. Ram and Paul Anbu K. (2014) examined the RMS use of Information professionals in India with a non-use rate of 61%. Sarrafzadeh and Hazeri (2014) found a non-use rate of over 40% in their sample of information professionals in Iran. No samples, even those of a similar size in this paper, have a rate of non-use as high as this sample of MLRs.

To further explore non-use, figure 10 separates Roman and non-Roman researchers. Non-Roman researchers tend to cluster around using Endnote and Zotero, as do Roman researchers. However, Roman MLRs also use a greater diversity of RMS. Regardless of Roman or Non-Roman, why does RMS use remain low?



Non-Users and Low RMS adoption rates

To understand the question posed, Non-Users were broken down into two groups, those who have never used RMS and those who quit using RMS. The "never used RMS" group (N=55), trend toward older demographics. In free text responses to "Why haven't you adopted an RMS," they stated feeling no need or being happy with their personal system for organizing sources. A fair number of participants (n=22) had never even heard that this kind of software exists. Others were not willing to adopt RMS because they were concerned about the learning curve. A few Non-Roman MLRs mentioned they did not want to risk using software that couldn't render the alphabets or ideograms that weren't Latin-based.



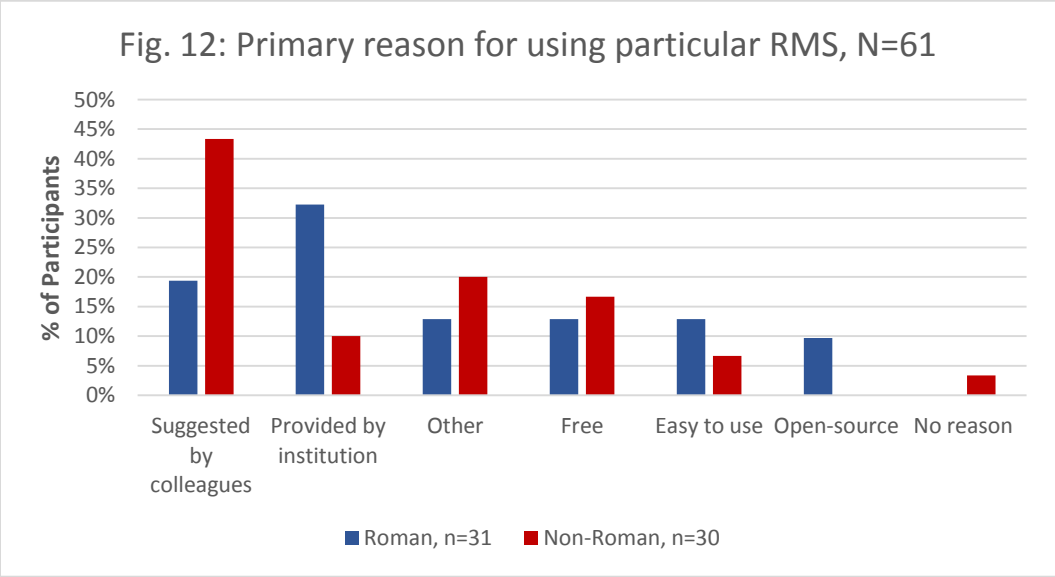
Figures 11a and 11b show that compared to other basic issues, multilingual ones are not the primary reason for avoiding RMS. Among non-users, only 25% were not aware that such software exists, and therefore could not asses why they would or would not use RMS. Only 5% of participants in the non-use group took multilingual issues into consideration. In contrast, 25% of the "quit using RMS" group said

multilingual issues factored into their decision to quit, with “poor performance” and “learning curve” mentioned most often.

The “quit using RMS” participants said it was easier to keep lists of sources in a word processor. Multilingual issues mentioned included the inability to handle non-Roman characters, and no ability to generate citations with some combination of vernacular script, transliterations, and/or translations. Roman MLRs disliked that some RMS would force English style capitalization, as well as the difficulty of inputting diacritics. After attempting to learn one RMS, there was little motivation to attempt learning another. Participants concluded that RMS were not saving them time or simplifying their workflow, two key enticements for adopting such software.

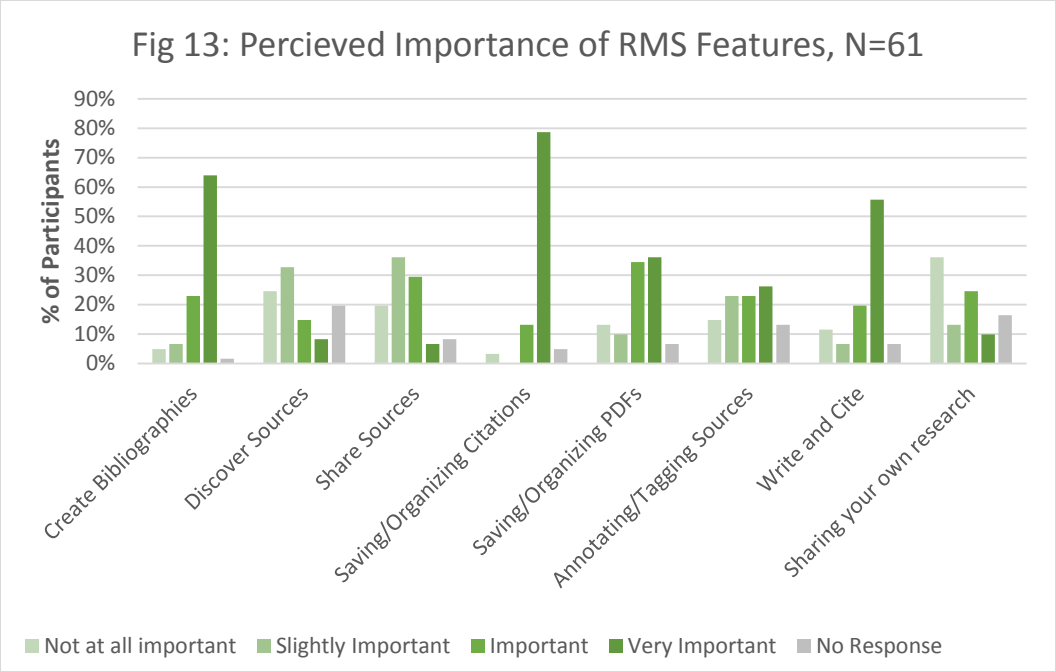
Users

In asking users their primary reason for selecting their RMS of choice, there are several noticeable differences between Roman and Non-Roman MLRs, highlighted in figure 12.



The primary factor for choosing an RMS for Roman MLRs appears to be access to institutional subscriptions, while Non-Roman MLRs rely on advice from colleagues. Because Roman MLRs face fewer logistical challenges in citing their sources, they may be fine with tools their institutions provide for free, regardless of their personal preferences. Non-Roman MLRs, on the other hand, having the additional burden of transliterating their sources, may feel their institutions’ “default options” are inadequate.

Figures 13 shows which features of RMS MLRs value most: Creating bibliographies, saving/organizing citations, and write and cite features.



27 participants provided open answer feedback to the question, "What improvements would you like to see in your RMS?" Thematically, the responses broke down according to: diacritics, handling non-Roman characters, handling right-to-left languages, and parallel fields to accommodate vernacular, transliteration, and translation of sources. One participant using Zotero noted that setting the language of the source in Zotero's metadata schema prevented the auto-capitalizing according to English rules. In contrast, a former Zotero user said they felt frustrated that the software would automatically capitalize their German sources. This user didn't know they needed to identify their source's language in the appropriate metadata field. This difference in familiarity with Zotero's features created a difference in perception of its capabilities.

Supporting Multilingual Researchers

While lack of support was never cited as a main reason for not adopting RMS, the barrier to entry for RMS among MLRs appears to be very high. Childress (2011) discusses the need for a strong support network of librarians to cover the diverse range of RMS available. However, librarians themselves also seem relatively unaware of multilingual issues based on the literature and the results of this survey. The survey did not explicitly ask if participants thought to ask the library for assistance with RMS, but Francese (2013) noted that consulting with a librarian was not common tactic when users had trouble with RMS. Additionally, many participants likely used older versions of these pieces of software. Their first impressions have remained despite advancements, better UI design, etc. The complex nature of multilingual citations means that librarians, in addition to being skilled at using and teaching RMS, need to have specialized knowledge of various languages that have their own conventions, capitalization, diacritics, characters and alphabets.

This creates a knowledge and technical literacy gap. Scholars do not use the technology because it is incapable of meeting their needs as MLRs. Therefore, they are not learning to use any RMS or how to participate in discussions about its development. When librarians engage and teach RMS to (potential)

users, it is important to teach users to be active participants in their research tools. Active participation provides an essential skill set for investigating software capabilities and to pose questions in a manner similar to how research questions are formed: developing a thesis, conducting preliminary research to asses gaps in knowledge, and drawing a conclusion that is based on objective facts instead of preconceptions. In the example regarding the auto-capitalization of non-English texts, if the researcher had investigated the problem, instead of assuming it was impossible to solve, they might have discovered a solution.

The Case of Juris-M

Throughout the figures in this article, one of the RMS listed is Juris-M. It is intentionally not grouped under "other" despite its low use count because it is specifically designed to manage multilingual citations. Only three participants were aware it existed, and only 1 participant was an active user. Juris-M is based off of Zotero, built directly from Zotero's code.

Juris-M's features parallel language fields, and the ability to adjust sources when the language metadata field is filled out. Typing "zh" "jp" or "ko" (Chinese, Japanese, and Korean, respectively) in the language metadata field subtly changes citations, adjusting the order of names automatically to show family names first, given names last. However, the initial learning curve remains problematic. Another issue is that Juris-M is marketed as a tool for law students and legal professionals, with extra metadata schema for legal citations. Juris-M was once called Multilingual Zotero, but its multilingual capacity is obfuscated by its new name and how it is promoted.

Limitations of Study

First and foremost, the variety of subgroups, including languages, scholarly disciplines, and RMS use limit the ability to pin point particular trends and habits of RMS users and non-users within the MLR community. As the size of the sub-samples shrinks they become statistically less relevant. A follow up survey, that captures a substantially larger sample, would provide a richer data set to draw more accurate conclusions.

A number of participants used multiple RMS, and the survey was not setup to account for this. A follow-up survey should account for a more detailed history of RMS use.

Conclusions

The design of RMS as it exists today creates a barrier to use. Different software products by multiple vendors fails to account for the needs MLRs have. Aside from the creation of parallel metadata fields, RMS should consider a design that allows diacritics to be typed easily. Mac users have this feature built into the Mac OS (Apple Inc. 2017) while Windows/PC users have to rely on workarounds as there is no efficient method for adding diacritics to Latin letters. Regardless of the operating system researchers use, RMS should incorporate a way to input diacritics easily as a fundamental part of their design to accommodate different Romanization standards.

MLRs who tried RMS in its early development found the software unable to accurately represent non-English information sources, or generate citations that met the guidelines of the journals they publish in. In turn, this lack of usability was likely reported out to colleagues giving RMS its poor reputation among MLRs.

While a large portion of MLRs in this sample reject RMS, those who adopted the technology appear to value its core functions: saving sources, making bibliographies, and write and cite functionality. What is not clear is if active RMS users are compromising the fidelity of the sources to use some of the automation and search features, or if they are compromising the ability to automatically output citations and bibliographies their publishers demand in order to take advantage of convenient, quick storage.

Some comments in the free text responses demonstrate technical literacy issues among MLR, suggesting that users do not thoroughly explore the capabilities of current RMS. This demonstrates a need for librarians to act as conduits for connecting researchers with appropriate software. However, area studies librarians specializing in geographic regions that use non-Roman scripts tend to be limited. Librarians must also incorporate components of information/technical literacy into teaching about RMS since they cannot be readily available to answer all questions, thus preparing users to confidently research which RMS will work best for how they approach their research.

To address the challenge associated with adapting RMS for multilingual needs and the technical literacy gap between MLRs and RMS, The author has created a guide for effectively using Juris-M (hyperlink and citation redacted). It was designed around the understanding that MLRs, despite being experienced researchers, may not feel comfortable with navigating RMS. Ultimately, even if people do not think of themselves as “multilingual,” access to research materials from around the world has become easier than ever, and research tools need to be designed with a baseline expectation that they are compatible with the research that happens across borders and languages.

Bibliography

Apple Inc. 2017. “How to Type Accents, Emoji, and Symbols on Your Mac.” *Apple Support*. Accessed March 15. https://support.apple.com/en-ca/HT201586.

Bennett, Frank G. 2013. *Citations, Out of the Box*.

Childress, Dawn. 2011. “Citation Tools in Academic Libraries: Best Practices for Reference and Instruction.” *Reference & User Services Quarterly* 51 (2): 143–152.

Francese, Enrico. 2013. “Usage of Reference Management Software at the University of Torino.” *JLIS.it* 4 (2): 145. doi:http://0-dx.doi.org.libraries.colorado.edu/10.4403/jlis.it-8679.

Harders, Harold. 2002. “Multilingual Bibliographies: Using And Extending the Babellib Package.” *TUGboat* 23 (3): 344–353.

Hicks, Alison, and Caroline Sinkinson. 2015. “Examining Mendeley: Designing Learning Opportunities for Digital Scholarship.” *Portal: Libraries and the Academy* 15 (3): 531–549. doi:10.1353/pla.2015.0035.

Hufflen, Jean-Michel. 2009. “Managing Languages Within Mlbibtex.” *TUGboat* 30 (1): 49–57.

(redacted)

Mead, Thomas L., and Donna R. Berryman. 2010. “Reference and PDF-Manager Software: Complexities, Support and Workflow.” *Brick and Click Libraries* 29 (4) (October): 388–393. doi:10.1080/02763869.2010.518928.

Melles, Anne, and Kathryn Unsworth. 2015. “Examining the Reference Management Practices of Humanities and Social Science Postgraduate Students and Academics.” *Australian Academic & Research Libraries* 46 (4) (December): 250–276. doi:10.1080/00048623.2015.1104790.

National Center for Educational Statistics. 2016. “Digest of Education Statistics.” https://nces.ed.gov/programs/digest/d16/tables/dt16_317.10.asp?current=yes.

Noguchi, Setsuko. 2009. “A Case Study of Web-Based Citation Management Tools with Japanese Materials and Japanese Databases.” *Journal of East Asian Libraries* (147) (February): 31–39.

Ram, Shri, and John Paul Anbu K. 2014. “The Use of Bibliographic Management Software by Indian Library and Information Science Professionals.” *Reference Services Review* 42 (3): 513–499.

Sarrafzadeh, Maryam, and Afsaneh Hazeri. 2014. “The Familiarity and Use of Reference Management Software by Lis Faculties in Iran.” *New Library World* 115 (11/12) (November): 558–570. doi:10.1108/NLW-02-2014-0018.

Wu, Ming-der, and Shih-chuan Chen. 2012. “How Graduate Students Perceive, Use, and Manage Electronic Resources.” *Aslib Proceedings* 64 (6) (November 23): 641–652. doi:10.1108/00012531211281779.