Using Data Mining for Citation Analysis

Philip B. White

This paper presents a new model for citation analysis, applying new methodological approaches in citation studies. These methods are demonstrated by an analysis of cited references from publications by the Geological Sciences faculty at the University of Colorado Boulder. The author made use of simple Python scripting, the Web of Science API, and OpenRefine to examine the most frequently cited journals and compare them to library holdings data to discover materials absent from the local collection. Of the more than 20,000 citations analyzed, 80 percent cited approximately 10 percent of all titles (412 journals). A notable finding was the heavy reliance of faculty members upon works between zero and two years of age. The streamlined model presented here removes the constraints of time and effort encountered by academic librarians interested in conducting citation analyses.

Introduction

Collecting and examining bibliometric data from works produced by academic library users provides librarians with solid information from which they can make data-driven decisions. Citation analysis is a bibliometric method used to identify patterns in scholars' publication habits such as how often an author or publication is cited or to identify networks in scholarly communication. Other applications of citation analysis include examining the literature of a specific field, analyses of scholar productivity and quality, studies of library patron groups, longitudinal studies of journal use, and assessments of library collections.[1] Studying the citation patterns of library user groups is a reliable way for academic librarians to develop a better understanding of their core users, which in turn guides them to a more sophisticated approach toward reference, instruction, and materials acquisition in their subject areas. Librarians began analyzing citations for collection development as early as the 1920s,[2] and the method has been widely used since the 1950s, when Eugene Garfield began his well-known work on the *Science Citation Index*.[3]

In academia, librarians build and maintain collections that support their institutions' researchers. Meeting such a responsibility can be complicated—librarians must not simply consider what disciplines are pursued; they must also gauge the vigor of those pursuits if they are to balance needs and build a collection that reflects varying levels of research activity. As a part of meeting this obligation, many academic librarians have applied citation analysis methods

*Philip B. White is Earth Sciences & Environment Librarian and Assistant Professor in the University Libraries at University of Colorado Boulder; email: philip.white@colorado.edu.

to identify information sources that are most critical to the success of faculty and students at their colleges and universities. A citation analysis of student or faculty works facilitates an assessment of the comprehensiveness of a library collection and informs the librarian's purchasing priorities. In simple terms, knowing the works library users cite, and with what frequency, helps identify gaps in the library's collection. If a faculty member or student cites a book or journal the library does not provide access to, the librarian can use that knowledge to determine if the item is suitable for acquisition.

Many studies conducted by library and information researchers have noted the utility of citation analysis and described the technique as unobtrusive and concrete.[4] Citation analysis also has advantages over other methods of collection assessment: citation data are readily available, reliably indicating use of the cited item. Citation analyses are time-intensive, however, particularly in disciplines where authors publish many articles a year. It is common for scientific journal articles to contain more than 50—or even 200—references in a given article. The high volume of works cited in the sciences makes citation analysis in these fields a challenge for librarians assessing a science collection.

Techniques developed in the fields of computer and data sciences offer a path toward fast and efficient analysis of large citation data sets. In a 2009 review of citation analysis studies, Allen Ashman suggests that future methodological approaches to citation studies will soon evolve to become deeper, more precise, and involve huge data sets—all incurring less effort by the researcher.[5] Ashman was correct; these methods have arrived and are waiting for librarians to take advantage of them.

Data mining, the practice of querying large databases to discover new information and create knowledge,[6] has the potential to open massive amounts of bibliometric information to librarians assessing their collections. The ability to access databases programmatically by way of application programming interfaces (APIs) represents an untapped resource in the field of citation analysis. An API is a set of protocols for building software applications and specifies how software components interact with each other.[7] APIs, in other words, allow computer applications to communicate with each other; the exchange of data and information is a frequent use of APIs. The methods sections of past citation studies often describe collecting data by obtaining book or article reference lists either one at a time or in small quantities—typically a manual process.[8] Use of an API to access bibliometric data reduces the amount of time spent obtaining citation data from weeks or months to minutes. Downloading citation data for individual articles or groups of articles from the Clarivate (formerly Thomson Reuters) Web of Science database would be time consuming and laborious at the scale of a few dozen articles or more. But librarians whose institutions subscribe to Clarivate's Web of Science Web Services Expanded (henceforth referred to as the Web of Science API) can access article reference lists over the internet in a programmatic fashion. Accessing large amounts of citation data using program code (in other words, a script) is an advantage because it allows for automated data extraction. A central aim of the research presented in this paper is to develop new methods of obtaining, manipulating, and analyzing a large amount of citation data in a relatively short amount of time.

The present study examines five years of citation data (January 2012 through December 2016) from publications by the Geological Sciences faculty at the University of Colorado Boulder (CUB). Earth and environmental sciences are important research fields at CUB. It is important that the CUB Libraries provide access to information resources for Geological Sci-

ences faculty and that its subject coverage in the discipline is broad and responsive enough to meet researchers' needs. The overarching goal of the study was to develop a fast and efficient method of obtaining citation data and to apply this new methodological approach to identify gaps in the University Libraries' earth sciences collection. The study pursued both collection assessment and methodological objectives. The specific objectives were:

- Discover key bibliometric trends of CUB Geological Sciences faculty publications including the publications in which the faculty's work most frequently appears, the most frequently cited publications, and the average age of works the faculty cite in their publications;
- Detect gaps in the earth sciences serials collection by identifying the most frequently cited works missing from the library's collection;
- Advance a streamlined, programmatic methodology for collecting citation data from the Web of Science database; and
- Develop an automated or semiautomated process for reconciling faculty citation data to the local library holdings data.

The research presented here advances citation studies by leveraging data retrieval and processing techniques from the field of data science. These advancements open up new possibilities to researchers seeking to conduct more expansive, comprehensive, and efficient citation analyses.

Literature Review

Citation Studies and Collection Assessment

While citation studies have a variety of applications, researchers evaluating library collections often categorize this line of research into two groups: (1) a global or worldwide analysis; and (2) a local or user group analysis.[9] In a global citation analysis, the researcher examines how often a particular publication or a selection of discipline-specific publications are cited without regard to the citing authors' institutional affiliations or geographic regions. An example of a global analysis is Melissa Rethlefsen and Lisa Wallis' 2007 study in which the authors reviewed three years of citation data from the *American Journal of Public Health* to identify the most frequently cited journal titles in the field of public health.[10] Global citation studies often make use of rankings lists such as *InCites Journal Citation Reports* (JCR) or similar products attempting to gauge a journal's "impact factor" by ranking how often authors cite a journal. Research by Alan Gale and Linda Day used JCR rankings of journals in several academic disciplines to determine if their library provided access to the most often-cited journals in those disciplines.[11] The authors noted that, while JCR rankings are indicative of the information needs of most authors in a particular field, the global approach might not accurately reflect the needs of researchers at their home institution.

Local citation studies focus on the citation habits of users affiliated with a particular institution, and researchers often consider this method a more reliable indicator of the library's collection development needs. Many researchers have argued that, because every library has a unique clientele, librarians should place greater priority on local users' citation habits—particularly when budget constraints demand a collection that is both lean as well as relevant to users' needs.[12] In a study of the usefulness of JCR rankings as indicators of local levels of serial use, Klaus Altmann and G.E. Gorman found that impact factors are not reliable data from which collection management decisions should be made.[13] Regression

analyses conducted by the authors to determine the relationship between impact factor and local serial use detected low regression coefficients, suggesting that the global approach to citation analysis is not a reliable predictor of local collection use. A contrasting study by Rick Ralston, Carole Gall, and Frances Brahmi correlated JCR Impact Factor of psychiatry journals with cited references of their institution's psychiatry faculty over a span of five years and found significant correlations between the two,[14] suggesting that JCR rankings are in fact useful for collection development. While debate exists over the quality of JCR rankings and similar metrics as a stand-in for local citation data, the literature firmly supports local user citation data as an indicator of collection relevance.[15] For these reasons, the author chose to make use of local users' citation data as a more definitive indicator of materials usage.

Local user citation analyses have proven beneficial to collection development needs, and librarians often use these studies to determine the extent of a library's subject coverage. These studies often reveal library collections that provide access to a majority of users' citations while also identifying some acquisition priorities. A local user study by Charlene Kellsey and Jennifer Knievel reviewed citations from 28 monographs published by humanities faculty members at their institution to determine the proportion of books cited in those works not owned by the library.[16] Their work identified deficiencies in the library collection. Out of 8,127 citations counted in faculty monographs, Kellsey and Knievel discovered that the library collection included 76 percent of the books cited. In another study conducted at the National Oceanic and Atmospheric Administration (NOAA) Miami Regional Library, Christie Wiley found that her library provided access to 74 percent of affiliated scientists' 2,156 cited references during a one-year period.[17] Others have detected much higher levels of coverage, particularly among serials. Susan Edwards and Lynn Jones analyzed a sample of citations from doctoral dissertations in the disciplines of education, psychology, and social welfare.[18] Edwards and Jones found that their library owned or provided access to greater than 97 percent of all journals cited in each discipline. While it is common for citation analyses to confirm the strength of a library collection, some researchers have identified subject areas with significant collection development needs. In a 2009 study, Jessica Kayongo and Clarence Helm found that their library provided access to only 41 percent of materials cited by Anthropology faculty at their institution, noting a heavy reliance on books as reference sources as a possible cause.[19]

While researchers have studied bibliometric trends in geological science, few citation analyses targeting collection development in the subject have taken place in recent years. Louise Zipp analyzed faculty and graduate student citations at the University of Iowa in 1996 and found that faculty most often cited the *Journal of Paleontology* followed by the *Geological Society of America Bulletin*.[20] A later global citation study by Zipp focused on identifying core journals in the subfield of Environmental Geology. Deploying an analysis of citation networks, Zipp identified 20 core journals, with *Environmental Geology* and *Ground Water* the top two in that subfield.[21] Salumi Helama researched age and material types of citations referenced in geology dissertations from the University of Helsinki.[22] Helama found that 65 percent of the dissertation references cited journals; the materials cited in the study were 14 to 17 years old on average. No recent study published in the library literature has brought the metrics analyzed by Zipp and Helama together with an assessment of local library holdings. The infrequency and relative lack of citation studies focusing on Geological Sciences adds significance to the present study.

The 80/20 Rule

When analyzing the proportion of materials cited by researchers to which the library provides access, librarians often refer to the 80/20 rule, an application of the Pareto Distribution to library serials first put forth by Richard Trueswell in the 1960s.[23] Applied to a citation analysis, the 80/20 rule implies that 80 percent of citations are attributable to 20 percent of the cited journals. Librarians have conducted analyses of the dispersal of citations to titles in efforts to identify core disciplinary collections. Hoffman and Doucette found that analyses of citation dispersion are a frequent component of citation studies.[24] Often, authors have labeled the top 20 percent of most frequently cited journals the core serial set for a specific discipline. Thomas Nisonger provides a comprehensive overview of the 80/20 rule and outlines ample supporting evidence to the rule in a 2008 literature review.[25]

Library researchers conducting citation analyses frequently both confirm and contest the 80/20 rule. Local user citation studies conducted by Keith Waugh and Margie Ruppel and another by Margaret Sylvia deviated from the 80/20 rule, with those studies reporting only 62 and 66 percent of citations coming from the top 20 percent of journal titles respectively.[26] More recently, studies published by Christie Wiley and Kimball et al. both found that 85 percent or more of citations analyzed in their research came from the top 20 percent, upholding—and even surpassing—the 80/20 rule.[27] While local differences in materials usage at different institutions limit application of the 80/20 to every case, the literature does suggest the majority of citations are to the minority of titles. The present study uses the 80/20 rule as a loose guide for identifying core serials.

Past Methodologies

There are various means of conducting citation analyses, yet researchers do not follow a standardized method of collecting and analyzing citation data. In a 2012 paper, Kristin Hoffmann and Lise Doucette reviewed the methodological approaches of 34 citation studies published between 2005 and 2010.[28] Hoffman and Doucette identify the typical variables analyzed in citation studies as types of resources cited, citation age, frequency of citation to journal titles, and (a check against) library holdings.[29] An important finding from Hoffman and Doucette's work is that citation studies are difficult to reproduce because authors inadequately describe their methods and rationale.

Of particular interest to the present study are the methods past researchers used to retrieve, refine, and analyze their citation data. Hoffmann and Doucette found that Web of Science was the most common data retrieval tool used by past researchers.[30] It is worth noting, however, that none of the studies reviewed made use of the Web of Science API. Researchers typically use the standard user interface of Web of Science and other library databases to conduct author searches followed by downloading authors' works or cited references—presumably one at a time or in batches. Only two studies have made use of programmatic scripts. A paper by Susann deVries, Robert Kelly, and Paula Storm and another by Johnathan Nabe and Andrea Imre both describe downloading unformatted citations from databases then applying a Perl script to parse elements of the citations into tabulated fields.[31] These studies took a step toward automation but stop short of using programmatic methods for data gathering.

Authors of citation studies have noted the time-consuming and difficult nature of analyzing citations from their institutions' science faculty due to the high volumes of published papers and cited references in those papers.[32] There are two common methods for mitigating

the difficulty of dealing with the volume of science citations. First, researchers often use citations from dissertations or theses as a proxy for cited references from faculty publications.[33] In a citation analysis of engineering dissertations, Madeline Kelly concedes, "While it would have been preferable to use faculty publications as the internal citation pool for this study, dissertations were easier to sample and thus more feasible given the time constraints of the project."[34] Some debate exists, however, over the use of graduate student work as a stand-in for faculty publications in citation analyses. In an often-cited 1996 study, Louise Zipp found positive correlations between thesis and dissertation citations and faculty citations at three universities, affirming the use of graduate student work as a stand-in[35] In contrast, a study by Yelena Pancheshnikov found that faculty publications cited a much broader variety of journal titles than citations from masters' theses.[36] She asserts that student theses are an unreliable substitute for faculty publications for citation analyses. A second approach toward dealing with very large amounts of citation data is to analyze a representative sample of a user group's citations to avoid a data set that is large and unmanageable.[37] In a study using citations from doctoral theses, Edwards and Jones sampled one out of every five citations from their initial data set.[38] While sampling and the use of student works as proxy are both commonly accepted, the new technical approaches presented in this paper brought greater efficiencies in data retrieval and refinement and made the need for proxy data and sampling unnecessary.

Data Science and Libraries

The gap between librarianship and data science is shrinking. In 2016, Frank Cervone published a review of the evolution of the field of data science and its application in Library and Information Science. He defines data science as "a transdisciplinary field that brings together statistics, computer science and information science and relies heavily on probability models, data mining and machine learning to help us understand and use the voluminous amount of data being created today."[39] He argues that the library community's contributions to the incorporation of data science into information studies are critical. In an article titled, "Teaching Librarians to Be Data Scientists," Christopher Erdmann advocates for the adoption of data science methods in librarianship and observes that these skills and techniques often allow for new partnerships between libraries and other data-intensive organizations.[40] He outlines the Data Scientist Training for Librarians (DST4L) initiative and discusses data science skills useful to librarians. Programs like DST4L and the Data Science and Visualization Institute for Librarians at North Carolina State University are introducing librarians to tools and methods for working with large data sets every year. In spite of these advancements, library researchers have yet to leverage data science techniques to enhance bibliometric and citation studies, making this study the first of its kind.

Methods

The CUB Libraries' earth sciences collection was selected for this research due to the libraries' desire to discover potential gaps in a collection that supports several disciplines. The Department of Geological Sciences, a primary user of the collection, has recently expanded its faculty, introducing new interdisciplinary research areas to the department. Five years (January 1, 2012 to December 31, 2016) of the department faculty's publication data were downloaded from CUB's local instance of Symplectic Elements, the university's platform for managing scholarly production.[41] Other researchers could obtain similar data

from Web of Science or InCites in the absence of Symplectic Elements. These data were stored in comma separated value (CSV) format, the fields of which included common bibliographic elements (title, date, and others) as well as additional information such as digital object identifiers (DOIs) and Web of Science accession numbers. Web of Science accession numbers indicate that a paper has been indexed by the Web of Science database; the accession number is a requirement when querying a document's reference data in the API. Out of 658 total publications by Geological Sciences faculty from 2012 to 2016, Elements provided Web of Science accession numbers for 431 of the papers. These 431 papers formed the sample of publications used in this study. Publications lacking Web of Science accession numbers were assumed to have appeared in publications not indexed by Web of Science. These materials consisted of monographs, conference proceedings, miscellaneous reports, Geology field guides, and some scientific journals. Cited references appearing in these items are likely of similar nature to those from publications used in the study. The author notes that cited reference data returned by the Web of Science are not limited to items indexed by the platform.

The Web of Science API does not have a graphical user interface. Rather, it is a Simple Object Access Protocol (SOAP) API, which relies on sending and receiving messages in Extensible Markup Language (XML). Authenticated users of the Web of Science API can send over the internet an XML query to the API, which will in turn provide a response message with the queried information in XML format. The XML request must contain the accession number for the publication queried, along with several other parameters that inform the API as to what type of search to conduct. One such parameter is a cited reference search, which returns the full reference list from the queried publication. Full documentation for the Web of Science API is available online.[42]

Next, the author developed a Python script for interacting with the Web of Science API. Python is a free and open source programming language with thousands of user-developed modules that permit automation of a wide variety of tasks.[43] Using the SUDS module to communicate with the API,[44] a script of Python code can automatically generate, send, and receive XML messages from a SOAP API. The script included all of the necessary parameters to execute a cited reference search and incorporated a "for loop" that generated a search for each accession number from the publications used in the study. The script accomplished the following tasks:

1. Opened the CSV file containing the accession numbers of all 431 publications used in the study;
2. Iterated through the accession number list, generating a new cited reference query for each in XML format;
3. Sent each query in sequence to the Web of Science API;
4. Received the XML response messages from the API for each query, which contained the full list of cited references of each publication;
5. Created a new file, appending and saving the XML response data from each query to that file.

In this instance, the script took approximately eight minutes to return the data. The script and instructions for its use are freely available online.[45] A list of all the variable fields returned by the cited reference query is available in table 1. The returned data represented citations to any material type—journals, books, government documents, or anything else.

TABLE 1
Variables Returned from Cited Reference Query

Variable	Description
queryId	ID number for individual query
docid	The cited work's WOS accession number
citedAuthor	First author of the cited work
timesCited	Number of times the cited work has been cited
year	Publication year of cited work
volume	Volume of publication the cited item appears in
page	Page of publication the cited item appears in
citedTitle	Title of cited work (typically article title)
citedWork	Title of publication (typically journal title)
recordsFound	Number of cited works in queried item's reference list
recordsSearched	Total records searched during query

The data returned from the Web of Science API-cited reference queries required some cleaning and standardization. The author transformed the cited reference data from XML to CSV using Microsoft Excel and then imported from CSV format into OpenRefine. OpenRefine is a free and open source software application used for data refinement tasks such as transformations, pattern detection, mass editing, and detection of inconsistencies.[46] OpenRefine provides a graphical user interface that lets users perform complex data transformations that otherwise would require advanced coding skills. An obstacle of working with this data set is that a Web of Science–cited reference query returns textual content formatted in the preferred style of the journal in which the publication appeared. For example, the *Journal of Geophysical Research: Atmospheres* could appear as such, or as "*J. Geoph. Res. Atmos.*," as "*Geophys. Res. A.*," or some other derivation. OpenRefine's semiautomated clustering tools allowed for standardization of journal titles from the citedWork field. This process allowed for easy tallying of total journal citation counts and provided a standardized titles list that could be more readily compared to library holdings data (for more on OpenRefine's clustering technology, see Verborgh and Wilde).[47] OpenRefine's text facets also quickly calculate how many instances of a particular value are present, which served to inform how many times each journal title had been cited.

Holdings data from the library's integrated library system helped verify which journal titles from the cited reference data were currently present in the library's serial holdings. OpenRefine's Reconcile Service was used to compare journal titles from the citedWork field against the serial holdings data. Reconciling journal titles from the two data sets was a semiautomated process. The Reconcile feature verified exact matches in journal title names automatically, but minor discrepancies between journal title punctuation and format between the two data sets meant that a portion of the reconciling process became a supervised procedure. To expedite the reconciling process, this study did not check cited works against the library's holdings that faculty cited fewer than four times over the five-year period. The rationale for this decision is that such a small amount of use does not justify adding an item to the collection.

Finally, the study used Microsoft Excel to calculate bibliometrics of the faculty publication list, the cited reference data, and the reconciled citedWork-holdings data. The metrics calculated include:

- Publications per year;
- Ranking of how often faculty members published in a journal;

- Total citations within those publications;
- Mean, median, and mode of citations per publication;
- Mean age of citation at time of citing;
- How often each journal was cited;
- Proportion of citations coming from the top 20 percent of serials;
- Proportion of journals cited available in the library's serial holdings.

Results and Discussion

Bibliometric Trends

The Geological Sciences faculty at CUB published prolifically during the five-year period of study, and the faculty's citation and publication patterns are quite similar. The publication data included works from 31 unique authors from the Geological Sciences Department. The faculty averaged 86 publications per year indexed by Web of Science. The 121 articles published in 2015 represented a high-water mark (see table 2). The journal most frequently published in was *Geophysical Research Letters*. The faculty's work appeared there 64 times—nearly double the number of times articles appeared in the second-most published-in journal, *Earth and Planetary Science Letters* (see table 3). Several high-impact, multidisciplinary journals (such as *Science* and *Nature*) are also present in the list of works most frequently published in. Table 4 shows the top 20 most frequently cited journals during the study period. Confirming its status as a preeminent geoscience publication, *Geophysical Research Letters* was the most cited title, with 1,074 citations. *Science* (933 citations) and *Nature* (734 citations) ranked numbers 2 and 3 for times cited, respectively. While there was some overlap between the most cited journals and the most published in journals, 8 of the top 20 most published-in journals were not among the top 20 most cited journals. This finding suggests that faculty do not always publish in the journals they

TABLE 2
Annual Summary of Geological Science Faculty Publications and Citations

Year	Publications	Total Citations in All Publications	Average Citations per Publication
2012	78	4,808	61.6
2013	65	3,621	55.7
2014	69	3,572	51.8
2015	121	6,324	52.3
2016	98	6,123	62.5
Total	**431**	**24,448**	**56.7**

TABLE 3
Rank of Journals Published in by Faculty 10 or More Times

Rank	Publication	Times Published In
1	*Geophysical Research Letters*	64
2	*Earth and Planetary Science Letters*	34
3	*Geology*	19
4	*Science*	17
5	*Journal of Geophysical Research. Earth Surface*	16
5	*Geosphere*	16
7	*Quaternary Science Reviews*	14
7	*American Mineralogist*	14
9	*Pure and Applied Geophysics*	13
10	*Journal of Geophysical Research. Space Physics*	11
10	*Geochimica et Cosmochimica Acta*	11
12	*Nature*	10
12	*Nature Geoscience*	10

TABLE 4
Top 20 Most Frequently Cited Journals

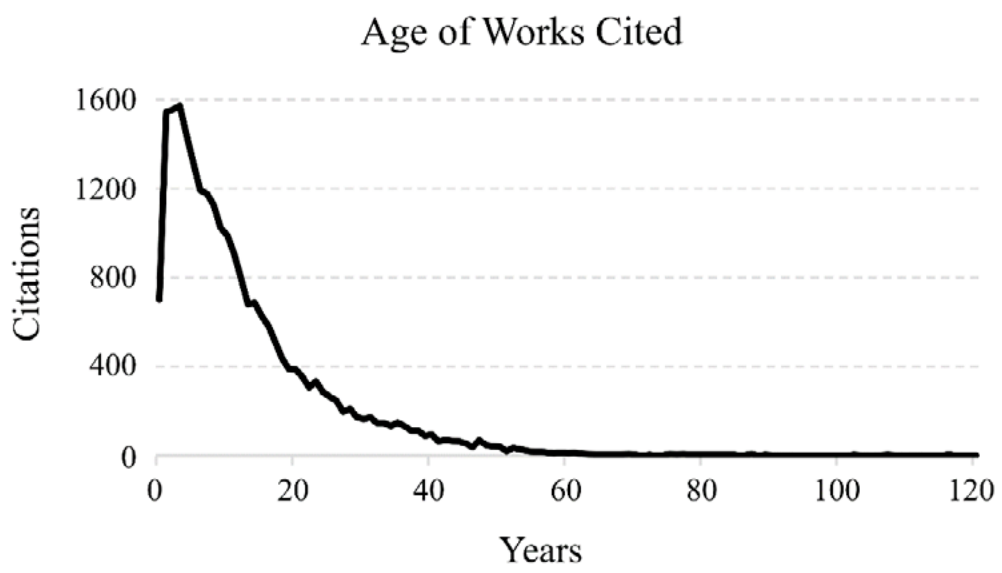
Rank	Publication	Times Cited	Times Published in by Faculty
1	*Geophysical Research Letters*	1,074	64
2	*Science*	933	17
3	*Nature*	734	10
4	*Earth and Planetary Science Letters*	638	34
5	*Geology*	559	19
6	*Journal of Geophysical Research. Solid Earth*	546	9
7	*Quaternary Science Reviews*	461	14
8	*Journal of Geophysical Research. Space Physics*	456	11
9	*Geochimica et Cosmochimica Acta*	412	11
10	*Space Science Reviews*	374	8
11	*Icarus*	350	5
12	*Journal of Geophysical Research. Planets*	321	5
13	*Geological Society of America Bulletin*	317	3
14	*Journal of Geophysical Research*	297	n/a*
15	*Journal of Geophysical Research. Atmospheres*	268	2
16	*American Mineralogist*	257	14
17	*Nature Geoscience*	237	10
18	*Geophysical Journal International*	233	1
19	*Proceedings of the National Academy of Sciences: PNAS*	214	3
20	*Journal of Geophysical Research. Earth Surface*	204	16

*Title split into multiple sections in 1978.

assign the most importance to—a reasonable assumption given that the most important journals typically have low acceptance rates. The Geological Sciences faculty overwhelmingly cite journals more than any other material resource. A review of the material types of the most cited works found that, among the 147 works cited 20 or more times, only two items were not standard serials: United States Geological Survey (USGS) Professional Papers and the Intergovernmental Panel on Climate Change (IPCC) assessment reports. These exceptions reflect both the importance of USGS reports as reference materials in the Geological Sciences and the faculty's focus on climate research.

Basic counts and age calculations of the works cited in the articles underscore the importance of new and emerging research to the Geological Sciences faculty. The reference lists for each of the 431 articles contained 24,448 cited references, averaging 56 citations per paper (see table 2). The median age of each cited reference at the time of the citing article's publication was nine years. The faculty, however, most often cited much more recently published articles. The citation age occurring most often was three years old (see figure 1), and approximately 22 percent of all the citations were three years old or less at time of publication. In fact, the faculty cited items aged zero or less 703 times—sometimes citing articles in press that were not due to be officially published for a year or more. The implication of this finding is that the faculty rely heavily upon works that are between zero and two years old, considering the lag between

FIGURE 1
Age of Cited Works at the Time of Citing Article's Publication

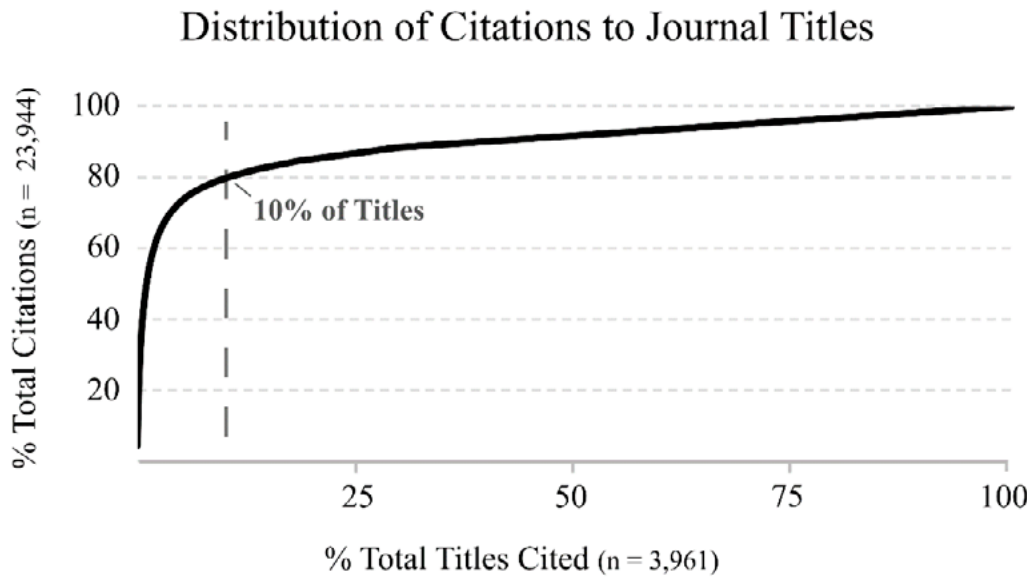


the time articles are written and the time they are published. Given the importance of recent research, the finding suggests that prohibitions on access to recently published items, such as the six-month to one-year embargoes publishers sometimes impose, may hinder scientists' work.

Analysis of dispersion of citations among the cited works' titles surpassed the 80/20 rule—that 80 percent of use comes from 20 percent of the titles.[48] The dispersal calculations in this study encompassed 23,944 citations, rather than the total number of 24,448 because approximately 2 percent of the citations returned by the API did not contain data in the cited work field (see *Study Limitations* below). In this analysis, 20 percent of the journal titles received 85 percent of the citations. These results parallel a dispersion analysis of a related subject field conducted by Kimball et al. at Texas A&M University.[49] A closer look at the dispersion of citations reveals that the Geological Sciences faculty rely the most on a set of serials constituting much less than the top 20 percent. While the papers reviewed for this study cited 3,961 unique titles, 80 percent of citations went to just 10 percent of titles (412 journals). In fact, nearly half of all citations (49.2 percent) went to only 1 percent of all titles cited (40 journals).

Figure 2 shows the relationship between number of citations and the percentage of titles cited, illustrating that a large portion of the cited references in this study cite a relatively small selection of all of the titles cited. This outcome indicates that, even though the Geological Sciences faculty cited a wide range of titles, they tended to rely the most on a small set of journals. If the 80 percent mark is indicative of the core serials collection for a subject, as many have proposed,[50] then this finding suggests that the core earth sciences serials at CUB are composed of approximately the top 10 percent of titles cited. The variance in results between the present study and other studies that have tested the 80/20 rule suggests that the dispersion of cited references to titles cited by researchers will vary at different institutions. While the 80/20 rule may be useful as a general rule of thumb, other librarians seeking to identify a core collection with precision would have to conduct similar studies locally. Future work comparing citations among similar faculty groups at different institutions (such as Geological Science Departments with comparable teaching and research foci) could identify a core earth sciences collection that could apply to many academic libraries.

FIGURE 2
Percentages of Total Citations vs. Total Titles Cited (Citations to the left of the vertical dashed line represent 80 percent of all citations, which cited only 10 percent of all titles.)



Library Holdings

The serial holdings offered by the library provided good coverage of the journals most often cited by the Geological Sciences faculty at the University of Colorado Boulder. Figure 3 depicts the proportions of cited materials that the library provided access to for items cited at least 20, 10, and 5 times. At the point where the journal titles reached 80 percent of citations (as discussed above), those journals received five citations during the five-year span of the study. At that level, the library provided access to 92 percent of all titles cited. The author expected high holdings coverage of the most important journals in the field. The earth sciences at CUB have had a dedicated branch library for 20 years and a subject librarian performing collection development for even longer.

Even though the library provided access to a high proportion of the most frequently cited materials, there were some exceptions. The library did not have a current subscription to 36 items that the faculty cited five times or more from 2012 through 2016. Table 5 shows works missing from the collection cited 10 times or more. Each item in table 5 is within the top 7 percent of most frequently cited titles. Notably, the fortieth most frequently cited work, *Quaternary Research*, is among the titles to which the library did not provide access. During the timespan of the study, the Geological Sciences faculty cited *Quaternary Research* 110 times—within the top 1.01 percent of most cited titles. Interestingly, concurrent to the writing of this paper, a faculty member requested that the library subscribe to *Jokull*, the fourth most cited serial in this study that is missing from the library's collection. While anecdotal, the request granted some cogency to the findings. Another noteworthy finding was that five out of the ten most frequently cited items not available from the library were Spanish language publications dealing with research in Patagonia. A clear coverage gap in this subject area likely indicates that these titles are important to one or two faculty members. Academic librarians seeking to replicate this study locally might similarly discover works of importance to the faculty miss-

FIGURE 3
Proportions of Cited Titles Included and Not Included in the Library's Collection at Three Levels of Citing Frequency

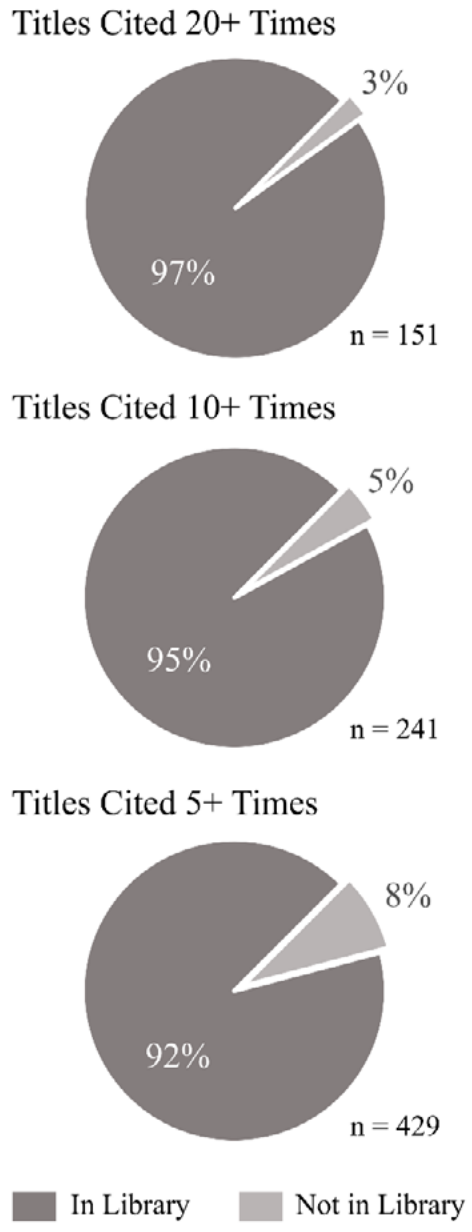


TABLE 5
Top Ten Most Cited Journals That Are Absent from the Library Collection

Rank	Publication	Times Cited
1	*Quaternary Research*	110
2	*Soil Science Society of America Journal*	28
3	*Anales Del Instituto de La Patagonia, Serie Ciencias Humanas*	27
4	*Jokull*	20
5	*Contra Viento y Marea. Arqueologia de Patagonia*	19
6	*Anales del Instituto de La Patagonia Serie Ciencias Sociales*	15
7	*Arctic*	13
8	*Sop Lando en el Viento. Actas de las III Jornadas de Arqueologia de la Patagonia*	11
9	*Photogrammetric Engineering and Remote Sensing*	10
9	*Arqueologia de Patagonia: Una Mirada Desde el Ultimo Confin*	10

ing from their collections. In the case of this study, the analysis produced clear collection development priorities. A logical next step in the process is consulting with Geological Sciences faculty and graduate students to determine which items among those missing from the collection are most important to their work.

Future research should seek to identify the means by which faculty obtain cited materials that are inaccessible to them via the library and delineate the effects of their habits of obtaining these items on the library. Presumably, researchers in need of materials that are embargoed—or not in the collection—obtain these items from interlibrary loan, personal subscriptions, or other means. Combining cited reference data with a faculty survey and interlibrary loan statistics would provide a more complete assessment of how material usage and obtainment affects the library. Unavailable items cited by faculty may be reflected in interlibrary loan (ILL) request records. Conversely, if frequently cited but unavailable items do not turn up in ILL records, then one could assume faculty obtain these materials by some other means. Anonymously surveying faculty might identify faculty members' means of obtaining these unavailable materials. A further question might entertain the economic effects of authorized or unauthorized obtainment of unavailable library materials on the library. Subsequent studies could investigate the rate of ILL requests of these materials and the presence or lack of increased ILL costs.

Methodological Implications

The methods advanced in this paper offer a compelling step forward in collecting citation data, and future work could expand and improve the techniques presented in this paper. Making use of the Web of Science API and a simple Python script significantly increased the speed with which an analysis of this scale can take place. Because of the increased efficiencies exhibited in the methods of this paper, librarians could potentially expand citation analyses from a project that takes place every few years to a regularly occurring assessment activity. Further, while this study may have been one of the largest citation studies ever in terms of total citations included, the method allows for expanding the scope of future projects to potentially huge volumes. With the burden of lengthy data collection periods removed, future researchers could significantly expand the scope of their citation analyses. A next step in this research would be for academic libraries to work together to examine several related disciplines together or conduct a cross-institution collection-use comparison. By conducting such analyses, academic libraries can make data-driven decisions about additions and removals from the collection. With this comprehensive collection-use data in hand, librarians involved in collection development can identify materials for both purchase and deselection—making better use of funds overall.

It is within reason that librarians could even work with data scientists and software developers to build automatically updating databases that obtain citations from their university's faculty and routinely cross-checks those data against the library catalog. A resource of this nature would be complex to develop, but it could provide near real-time collection assessment information. This study made use of both automated and semiautomated data retrieval and refinement techniques. Future work should look to expand on the use of data science methods in evaluating library collections and could aim to fully automate many of these procedures.

Study Limitations

Although citation analysis is a proven method of obtaining information about library materials usage, the method has some limitations. While a cited reference suggests that an author used an item, a list of citations does not necessarily include all materials used by an author. Researchers do not always cite every item they read or download from a library database. It is also possible that reference lists and bibliographies underrepresent the use of a class of materials, such as reference works. The opposite may also be true in that some materials cited may have an inflated value. In a paper detailing citation analysis methods, Margaret Sylvia discusses how authors sometimes cite materials that are of marginal importance to their research, cite another author to curry favor, or cite themselves or friends to increase their number of citations.[51] Sylvia also mentions that authors may be more likely to cite materials the library provides access to, which could skew the data toward including a disproportionate number of materials in the library collection. Citations could also be incomplete or inaccurate. Peggy Johnson points out that cited references have an inherent time lag, which can obscure the importance of newer journals and changes within a discipline.[52] The research presented in this paper focuses on materials used by faculty, but the collection is used by more than just faculty members. A full assessment of the completeness and relevance of a local collection in all aspects should seek to encompass metrics of graduate student, undergraduate, and, in some cases, public use.

There were some complicating factors and limitations to the work presented in this paper. First, the cited reference data returned from the Web of Science API contained some minor

inconsistencies. Occasionally, some data were missing. Approximately 2 percent of the citations were missing the journal title from the citedWork field. It was up to the author to decide if these omissions were intentional or accidental; they were ultimately not included in the reconciliation process. Many of the inconsistencies may boil down to unintentional mistakes or omissions in the citation list by the original author, which are largely unavoidable. The high volume of data used in this study, however, likely mitigated the effect of such inconsistencies. The varying styles that journals use to format bibliographic references was another complicating factor. The most time-consuming aspect of the project was standardizing names of cited works—among both the citation data and the library's holdings data.

Reproducing this study in the social sciences or humanities would be difficult. Journals and books in these disciplines often use the endnotes style of bibliographies. Endnotes typically list individual works repeatedly in the bibliography for each time it is cited, which would require deduplication of the cited reference data from each individual publication to avoid artificially inflating the results. The author also notes that the methods used in this study work best when analyzing disciplines that rely heavily on academic journals. Future researchers seeking to apply similar methods to citation analysis of humanities fields may find that Web of Science lacks comprehensive coverage of nonjournal materials. Future research devising methods of streamlining citation analyses for books is needed.

While the Web of Science API allows for great speed and efficiency, future researchers aiming to take advantage of these methods must have institutional subscriptions to Web of Sciences Web Services Expanded to access cited reference data following the techniques presented here. The author is unaware of any other web service that provides complete cited reference data for individual papers. In addition, the use of Web of Science as the data source excludes those items not indexed by Web of Science, which could result in an underrepresentation of new journals, journals increasing in importance, and journals from emerging fields.

Conclusion

The study presented in this paper approached collection assessment from a data science perspective, bringing new methods to the time-honored practice of citation analysis. The study successfully met its dual objectives of assessing the comprehensiveness of the earth sciences collection and advancing technological approaches to citation studies. The research identified core earth sciences serials, found gaps in the local collection, and developed new methodological techniques for citation analysis. The results of this study will allow the author to improve the earth sciences collection and develop a more informed plan for reference and instruction services.

The results of the analysis of Geological Sciences faculty citations produced useful insights into how the earth sciences collection is used. The library provided access to a high proportion of the most frequently cited journals. Yet, even with good coverage, the detection of important serials missing from the collection proved the value of this research and provided the author with information that will improve the library's services. The extent to which the results exceeded the 80/20 rule was an interesting surprise; the fact that 80 percent of citations referenced only the top 10 percent of journals (and nearly 50 percent to just the top 1 percent) emphasized the faculty's reliance on a relatively small set of publications. Obtaining and maintaining full access to these core publications should be a priority moving forward.

This study produced several implications for library practice. First, the results of this work provided a needed update to the literature regarding earth sciences collections and faculty use. Subject librarians developing and maintaining earth science collections may benefit from this study's identification of highly cited and published-in serials. The most relied-upon journals identified in this research have changed when compared to the last study of this type within the Geological Sciences, perhaps reflecting changes in the discipline.[53] An unexpected finding was the faculty's heavy reliance on works published within the preceding three years, with prepublication works cited more than 700 times during the five-year span. This finding emphasizes the importance of recently published works and preprints to faculty research. Further work analyzing citations to prepublication materials across the STEM fields could help librarians determine the extent to which publisher embargoes affect knowledge dissemination in the sciences. The work presented here also provides an efficient model for how collection development practitioners can identify high and low-use library materials, allowing for data-driven selection and deselection of serials.

Advancements in data collection and refining methods allowed for great speed and efficiency in conducting this research project. The author hopes that other librarians will adopt and expand these methods to learn about their faculty constituents' research needs and to identify gaps in their library collections. The ease with which the analysis took place offers a compelling reason for other librarians interested in collection assessment to make use of the new model presented here. Assessments of this type ultimately lead to better use of funds and improved understanding of the collection. Obtaining information for making data-driven decisions about library collections has never been easier, and librarians have an opportunity to take full advantage of the new techniques available to them.

Acknowledgements

The author thanks Jack Maness and Yem Fong for reviewing early drafts of the article. The author also wishes to acknowledge the helpful criticism of three anonymous reviewers.

Notes

1. Allen B. Ashman, "An Examination of the Research Objectives of Recent Citation Analysis Studies," *Collection Management* 34, no. 2 (Apr. 1, 2009): 112–28, doi:10.1080/01462670902725885.

2. P.L.K. Gross and E.M. Gross, "College Libraries and Chemical Education," *Science* (New Series) 66, no. 1713 (1927): 385–89.

3. Tony Cawkell and Eugene Garfield, "Institute for Scientific Information," *Information Services and Use* 21, no. 2 (2001): 79–86.

4. Kristin Hoffmann and Lise Doucette, "A Review of Citation Analysis Methodologies for Collection Management," *College & Research Libraries* 73, no. 4 (July 1, 2012): 321–35, doi:10.5860/crl-254; Jennifer Knievel and Charlene Kellsey, "Citation Analysis for Collection Development: A Comparative Study of Eight Humanities Fields," *The Library Quarterly: Information, Community, Policy* (Apr. 1, 2005), 142–68, doi:10.1086/431331.

5. Ashman, "An Examination of the Research Objectives of Recent Citation Analysis Studies."

6. Sudha Ram, "Data Mining," *Computer Sciences*, ed. Roger R. Flynn (New York: Macmillan Reference USA, 2002).

7. Vangie Beal, "What Is API: Application Program Interface? Webopedia," webopedia.com, available online at www.webopedia.com/TERM/A/API.html [accessed 16 June 2017].

8. Hoffmann and Doucette, "A Review of Citation Analysis Methodologies for Collection Management."

9. Peggy Johnson, *Fundamentals of Collection Development and Management, 3rd Rev. Ed.*, 3rd ed. (Chicago: American Library Association, 2013).

10. Melissa L. Rethlefsen and Lisa C. Wallis, "Public Health Citation Patterns: An Analysis of the American

Journal of Public Health, 2003–2005," *Journal of the Medical Library Association: JMLA* 95, no. 4 (Oct. 2007): 408–15, doi:10.3163/1536-5050.95.4.408.

11. Alan Gale and Linda Day, "Characterizing Journal Access at a Canadian University Using the Journal Citation Reports Database," *Partnership: The Canadian Journal of Library and Information Practice and Research* 6, no. 1 (2011), available online at http://search.proquest.com.colorado.idm.oclc.org/lisa/docview/907926259/3622C3A9C7B545CEPQ/2 [accessed 19 June 2017].

12. Maurice B. Line, "Use of Citation Data for Periodicals Control in Libraries: A Response to Broadus," *College & Research Libraries* 46, no. 1 (1985): 36–37.

13. Klaus G. Altmann and G.E. Gorman, "Can Impact Factors Substitute for the Results of Local Use Studies? Findings from an Australian Case Study," *Collection Building* 18, no. 2 (June 1, 1999): 90–94, doi:10.1108/01604959910265878.

14. Rick Ralston, Carole Gall, and Frances A. Brahmi, "Do Local Citation Patterns Support Use of the Impact Factor for Collection Development?" *Journal of the Medical Library Association: JMLA* 96, no. 4 (Oct. 2008): 374–78, doi:10.3163/1536-5050.96.4.014.

15. Hoffmann and Doucette, "A Review of Citation Analysis Methodologies for Collection Management."

16. Charlene Kellsey and Jennifer Knievel, "Overlap between Humanities Faculty Citation and Library Monograph Collections, 2004–2009," *College & Research Libraries* 73, no. 6 (Nov. 2012): 569–83.

17. Christie A. Wiley, "Using Citation Analysis to Explore the Collection Needs of Atmospheric Scientists/Researchers Affiliated with the Atlantic Oceanographic Meteorological Laboratory," *Library Collections, Acquisitions, & Technical Services* 38, no. 3/4 (July 2014): 82–91, doi:10.1080/14649055.2015.1080509.

18. Susan Edwards and Lynn Jones, "Assessing the Fitness of an Academic Library for Doctoral Research," *Evidence Based Library & Information Practice* 9, no. 2 (Apr. 2014): 4–15.

19. Jessica Kayongo and Clarence Helm, "Citation Patterns of the Faculty of the Anthropology Department at the University of Notre Dame," *Behavioral & Social Sciences Librarian* 28, no. 3 (Sept. 4, 2009): 87–99, doi:10.1080/01639260903089040.

20. Louise S. Zipp, "Thesis and Dissertation Citations as Indicators of Faculty Research Use of University Library Journal Collections," *Library Resources & Technical Services* 40, no. 4 (1996): 335–42, doi:10.5860/lrts.40n4.335.

21. Louise S. Zipp, "Core Serial Titles in an Interdisciplinary Field: The Case of Environmental Geology," *Library Resources & Technical Services* 43, no. 1 (Jan. 1999): 28–36.

22. Samuli Helama, "A Review of Citation Patterns in Doctoral Dissertations at the Department of Geology, University of Helsinki, Finland, since 1896," *Science & Technology Libraries* 31, no. 2 (Apr. 1, 2012): 180–89, doi:10.1080/0194262X.2012.676870.

23. Thomas E. Nisonger, "The 80/20 Rule and Core Journals," *Serials Librarian* 55, no. 1–2 (July 3, 2008): 62–84, doi:10.1080/03615260801970774.

24. Hoffmann and Doucette, "A Review of Citation Analysis Methodologies for Collection Management."

25. Nisonger, "The 80/20 Rule and Core Journals."

26. C. Keith Waugh and Margie Ruppel, "Citation Analysis of Dissertation, Thesis, and Research Paper References in Workforce Education and Development," *Journal of Academic Librarianship* 30, no. 4 (July 2004): 276–84, doi:10.1016/j.acalib.2004.04.003; Margaret J. Sylvia, "Citation Analysis as an Unobtrusive Method for Journal Collection Evaluation Using Psychology Student Research Bibliographies," *Collection Building* 17, no. 1 (Mar. 1998): 20–28, doi:10.1108/01604959810368965.

27. Wiley, "Using Citation Analysis"; Rusty Kimball et al., "A Citation Analysis of Atmospheric Science Publications by Faculty at Texas A&M University," *College & Research Libraries* 74, no. 4 (2013): 356–67.

28. Hoffmann and Doucette, "A Review of Citation Analysis Methodologies for Collection Management."

29. Ibid.

30. Ibid.

31. Susann deVries, Robert Kelly, and Paula M. Storm, "Moving beyond Citation Analysis: How Surveys and Interviews Enhance, Enrich, and Expand Your Research Findings," *College & Research Libraries* 71, no. 5 (2010): 456–466; Jonathan Nabe and Andrea Imre, "Dissertation Citations in Organismal Biology at Southern Illinois University at Carbondale: Implications for Collection Development" (2008), available online at https://works.bepress.com/jonathan_nabe/4/ [accessed 20 June 2017].

32. Edwards and Jones, "Assessing the Fitness of an Academic Library for Doctoral Research"; Kristina Romić and Gornaka Mitrović, "Using Citation Checking of Ph.D. Dissertation References as a Tool for Evaluating Library Collections of the National and University Library in Zagreb," *Libraries in the Digital Age (LIDA) Proceedings* 13, no. 0 (June 16, 2014), available online at http://ozk.unizd.hr/proceedings/index.php/lida/article/view/134 [accessed 17 February 2017]; deVries, Kelly, and Storm, "Moving beyond Citation Analysis."

33. Waugh and Ruppel, "Citation Analysis of Dissertation, Thesis, and Research Paper References"; Nabe and

Imre, "Dissertation Citations in Organismal Biology"; Pali U. Kuruppu and Debra C. Moore, "Information Use by PhD Students in Agriculture and Biology: A Dissertation Citation Analysis," *portal: Libraries and the Academy* 8, no. 4 (Oct. 11, 2008): 387–405, doi:10.1353/pla.0.0024; M. Kelly, "Citation Patterns of Engineering, Statistics, and Computer Science Researchers: An Internal and External Citation Analysis across Multiple Engineering Subfields," *College & Research Libraries* 76, no. 7 (Nov. 1, 2015): 859–82, doi:10.5860/crl.76.7.859; Edwards and Jones, "Assessing the Fitness of an Academic Library for Doctoral Research"; Maria Bernardete Martins Alves et al., "Correlation Between Information Needs and the Library Collection: A Citation Analysis Study of Doctoral Theses at Universidade Federal De Santa Catarina Library," *IATUL Annual Conference Proceedings*, no. 35 (June 2014): 1–10.

34. Kelly, "Citation Patterns of Engineering, Statistics, and Computer Science Researchers."

35. Zipp, "Thesis and Dissertation Citations as Indicators of Faculty Research Use of University Library Journal Collections."

36. Yelena Pancheshnikov, "A Comparison of Literature Citations in Faculty Publications and Student Theses as Indicators of Collection Use and a Background for Collection Management at a University Library," *Journal of Academic Librarianship* 33, no. 6 (Dec. 1, 2007): 674–83, doi:10.1016/j.acalib.2007.09.011.

37. Hoffmann and Doucette, "A Review of Citation Analysis Methodologies for Collection Management."

38. Edwards and Jones, "Assessing the Fitness of an Academic Library for Doctoral Research."

39. H. Frank Cervone, "Informatics and Data Science: An Overview for the Information Professional," *Digital Library Perspectives; Bingley* 32, no. 1 (2016): 7–10.

40. Christopher Erdmann, "Teaching Librarians to Be Data Scientists," *Information Outlook (Online); Alexandria* 18, no. 3 (June 2014): 21–24.

41. CU Boulder Elements (CUBE), available online at www.colorado.edu/fis/CUBE [accessed 9 November 2018].

42. Web of Science Web Services Expanded documentation is available online at http://ipscience-help.thomsonreuters.com/wosWebServicesExpanded [accessed 20 June 2017].

43. "Welcome to Python.Org," Python.org, available online at https://www.python.org/about/ [accessed 21 June 2017].

44. Documentation for the Python SUDS module is available at https://bitbucket.org/jurko/suds/wiki/Original%20Documentation [accessed 22 June 2017].

45. The Python script used for this study is available online at https://github.com/outpw/WOKapiscripts [accessed 1 September 2017].

46. Ruben Verborgh and Max De Wilde, *Using OpenRefine: The Essential OpenRefine Guide That Takes You from Data Analysis and Error Fixing to Linking Your Dataset to the Web* (Birmingham: Packt Publishing 2013).

47. Ibid., 52.

48. Nisonger, "The 80/20 Rule and Core Journals."

49. Kimball et al., "A Citation Analysis of Atmospheric Science Publications by Faculty at Texas A&M University."

50. Nisonger, "The 80/20 Rule and Core Journals."

51. Sylvia, "Citation Analysis as an Unobtrusive Method for Journal Collection Evaluation."

52. Johnson, *Fundamentals of Collection Development and Management, 3rd Rev. Ed.*

53. Zipp, "Thesis and Dissertation Citations as Indicators of Faculty Research Use of University Library Journal Collections."