ISSN 2161-3974

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

Introducing Reproducibility to Citation Analysis: a Case Study in the Earth Sciences

Samantha Teplitzky[1], Wynn Tranfield[2], Mea Warren[3], and Philip White[4]

[1] University of California, Berkeley, CA, USA
[2] University of California, Los Angeles, CA, USA
[3] University of Houston, Houston, TX, USA
[4] University of Colorado, Boulder, CO, USA

Abstract

Objectives:

- Replicate methods from a 2019 study of Earth Science researcher citation practices.
- Calculate programmatically whether researchers in Earth Science rely on a smaller subset of literature than estimated by the 80/20 rule.
- Determine whether these reproducible citation analysis methods can be used to analyze open access uptake.

Methods: Replicated methods of a prior citation study provide an updated transparent, reproducible citation analysis protocol that can be replicated with Jupyter Notebooks.

Correspondence: Samantha Teplitzky: steplitz@berkeley.edu

Received: October 23, 2020 **Accepted**: March 14, 2021 **Published**: May 13, 2021

Data Availability: Data and code are available at the following GitHub repository and will be preserved on Zenodo upon publication: https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project

Disclosures: The authors report no conflict of interest.

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Abstract Continued

Results: This study replicated the prior citation study's conclusions, and also adapted the author's methods to analyze the citation practices of Earth Scientists at four institutions. We found that 80% of the citations could be accounted for by only 7.88% of journals, a key metric to help identify a core collection of titles in this discipline. We then demonstrated programmatically that 36% of these cited references were available as open access.

Conclusions: Jupyter Notebooks are a viable platform for disseminating replicable processes for citation analysis. A completely open methodology is emerging and we consider this a step forward. Adherence to the 80/20 rule aligned with institutional research output, but citation preferences are evident. Reproducible citation analysis methods may be used to analyze open access uptake, however, results are inconclusive. It is difficult to determine whether an article was open access at the time of citation, or became open access after an embargo.

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Introduction

Librarians at research institutions have a keen interest in the citation activity of their scholars. The journals researchers publish in and the articles they cite give librarians a clear understanding of their liaison departments. Librarians can ensure they are providing the resources researchers need, and on the same coin, determine which resources may be irrelevant. A frequent analysis is ideal since departmental focus shifts with incoming and outgoing faculty members, but citation analysis has been a time-consuming endeavor. As a result, it is seldom tackled at a departmental level.

Existing citation analysis studies often present notoriously opaque methodologies. Applying programmatic methods similar to those introduced by Philip B. White in his 2019 paper, "Using Data Mining for Citation Analysis," our study introduces an analysis that aligns with the principles of open science. We provide a transparent, reproducible citation analysis protocol that can be replicated by librarians with basic coding skills. Four Earth Science librarians sought to expand on White's methodology and replicate the analysis at their distinct institutions. Results from several institutions provide the additional benefits of comparative analysis, and promotes collaboration among subject librarians. In sum, librarians represented Earth Science departments from the University of Colorado, Boulder (Boulder), the University of California, Berkeley (Berkeley), the University of California, Los Angeles (UCLA), and the University of Houston (Houston).

The study focuses on reproducible methods for citation analysis at the institutional level. In order to get a sense of how generalizable publishing practices are in the Earth Sciences, and how generalizable our process is for other institutions and disciplines, we turn to broad institutional rankings. The four institutions accounted for about 8% of Earth Science research articles from US institutions in the Web of Science database during this period. Boulder, Berkeley, and UCLA were within the top 25 institutions by articles published. Boulder represented the largest proportion (3.5%). Berkeley (2.3%) and UCLA (1.8%) are similarly sized, and Houston (.61%) represented the smallest in comparison.[1] Having four institutions with different levels of research output allowed us to test the validity of these methods.

A pre-registration was submitted to Open Science Framework[2] to set firm guidelines for data gathering and analysis. Foremost, we sought to replicate White's (2019) methods to programmatically assess the publication and citation practices of Earth Science researchers at other universities. We then analyzed whether or not researchers in Earth Science rely on a smaller subset of literature than estimated by the 80/20 rule, which states that 80% of publications cite just 20% of the literature (Nisonger 2008). Finally, we applied our methods to analyze open access uptake and citation of preprints.

1 Collected from Web of Science on 8/31/20. 214,100 articles were indexed as Earth Science or Geology for US institutions, 2010-2019.

2 https://osf.io/u49zv

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Literature Review

Previous citation studies, starting with seminal works in the field of Earth Sciences and concluding with an examination of current and emerging computational options, inform our work. In our cornerstone study, White (2019) notes the lack of local analysis targeting the geological sciences. Furthermore, existing studies in the geological sciences emphasize thesis and dissertation analysis as opposed to journal publication analysis (Zipp 1996; Walcott 1992; Helama 2012). These works were likely selected over faculty publications because a comprehensive institutional list offered a starting point for the search. Exceptions to this focus include Frohlich and Resler's (2001) analysis of the University of Texas Institute for Geophysics publications, which expanded beyond theses to include faculty publications.

No citation analysis is complete without a nod to the 80/20 rule. In a 2008 literature review, Nisonger describes the librarians' rule of thumb that eighty percent collection use can be attributed to just twenty percent of the collection. This Pareto distribution describes the "vital few" and "trivial many" applicable to myriad circumstances. In a library context, this phenomenon has been applied to book circulation wherein eighty percent of a library's circulation can be attributed to the twenty percent most used books in the library (Trueswell 1969). Others have observed the 80/20 rule in studies of cited references, where 80% of references cite just twenty percent of all of the titles cited (Fleming and Kilgour 1964; Eckman 1988; Sennyey, Ellern, and Newsome 2002). The 80/20 usage pattern is most useful for librarian researchers seeking to identify parts of a core collection. Examples and aspects of the 80/20 rule are abundant in Nisonger's (2008) deep dive into the phenomenon. The present study, as an analysis of citations, examines the 80/20 rule as it pertains to the proportion of citations to all titles cited in the study sample.

Definitions of reproducibility and replicability vary by discipline. In this study, we followed the designations of Whitaker (2017) and Clyburne-Sherin (2020) who define reproducibility as arriving at the same results using the same methods and the same data in contrast to replicability which applies the same methods with different data. Citation analyses are seldom standardized and often difficult to either reproduce or replicate. Hoffman and Doucette (2012) found Clarivate's Web of Science to be the most frequently used citation retrieval tool. This tool was used by Antelman (2004) in a multi-disciplinary analysis of the impact of Open Access articles, and again by Arendt, Peacemaker, and Miller in their 2019 attempt to replicate the study. Searches were executed manually and downloaded in batches, helping to earn citation studies a reputation for being exceptionally time-consuming. In the case of Arendt, Peacemaker, and Miller's (2019) replication, the Web of Science search engine also became the initial, immutable source of error in replication. Sampling tactics, such as selecting a subset of articles to represent the whole body of literature are often used to make spreadsheet anchored studies more manageable.

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Among published citation analysis studies, the most common tool used in post-collection analysis is a spreadsheet (Hoffmann and Doucette 2012). Though effective tools, spreadsheets' lack of automation when cleaning and filtering data makes it difficult for researchers to replicate results without highly detailed step-by-step instructions. Automating unformatted citations using programmatic scripts was first utilized by Nabe and Imre (2008) to parse dissertation citations. This was followed by deVries, Kelly, and Storm (2010), who automated parsing citation data into spreadsheet columns, but manually completed the "arduous process" of interpreting the parsed citations. The movement toward automation was taken a step further by White's (2019) case for using the Web of Science API—a tool available to institutional subscribers—to automate downloads and parsing of citation elements. In White (2019), OpenRefine's semi-automated clustering tools were used to standardize elements such as journal tiles and provide relevant counts, and OpenRefine's Reconcile Service compared the list to library holdings. Sterman and Clark (2017) use RSS feeds from multiple databases, including subscriptions and open access resources, to automate their initial citation gathering. They are, however, unclear about the exact databases they selected and the method of data cleaning. White (2019) also pointed out that using the Web of Science limits analysis to items indexed in the database, possibly leading to an underrepresentation of new or emerging journals.

Open Access (OA) has changed the publishing landscape over the past 20 years and adoption rates are of great interest in citation analysis studies; however, OA adoption by the Earth Sciences community has not been well-studied. Factors influencing uptake include institutional or grant mandates, such as National Science Foundation's Public Access Mandate (NSF n.d.). Complicating any analysis is the definition of OA itself and the many gradients of OA, well described by Piwowar et al. (2018). It is difficult for researchers completing a citation analysis to determine whether an article was published in Gold, Hybrid, Delayed, or Bronze OA formats since the format can change over time. Several metastudies consider OA uptake rates overall and offer useful benchmarks for Earth Science, though the boundaries of Earth Science as a discipline vary, as do the time periods covered and definitions of OA types. Piwowar et al. (2018) found that just over 40% of all DOI-assigned journal articles published in the Earth Sciences between 2009-2015 were available at some level of OA. Martín-Martín et al. (2018) estimated about 30% of Earth Science articles and reviews with a DOI published in 2009 and 2014 achieved some level of OA. Archambault et al. (2014) report just over 50% of Earth and Environmental Science articles as OA between 2011-2013, with the majority categorized as "Other OA" which includes hybrid and embargoed articles as well as those hosted at commercial sites like Researchgate.

At this juncture, a corpus of open and replicable methodology for citation analysis is emerging. Open Science Framework (OSF) has created a space for collaborative, pre-registration that has been embraced by myriad disciplines, including information professionals. King et al. (2016) used the OSF to document a project beginning in 2016 to examine the self-citation patterns of academics. Their robust documentation, including Stata code, favors reproducibility. Arendt, Peacemaker,

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

and Miller (2019) also used OSF to document their replication of Antleman's (2004) Open Access research impact analysis, providing a clear methodological checklist and accompanying script. They registered the study with OSF at its inception, making their initial research questions, hypothesis, and sampling plan transparent. The use of preregistration in citation analysis, not unlike registration for systematic reviews in health sciences, introduces rigor and necessitates foresight.

With the increased intersection of librarianship and data science, particularly in large-scale citation analysis, librarians are turning to collaborative data science tools. Many of these tools have already been adopted in education and scientific computing environments (Perkel 2018; Borowski et al. 2020; Stoudt, Vasquez, and Martinez 2020). GitHub allows asynchronous work, with a strong emphasis on writing distinct code. Literate computing, facilitated by Jupyter Notebooks or Google Colaboratory, encourages users to produce a document that stores the code, computational results, and observations in one place. As creators Perez and Granger (2015) note, humans process the world through narratives. Narrativizing a computational maneuver may lead to loss in translation when it is of the utmost importance that these same computational narratives are reproducible. Using Jupyter Notebooks enables collaboration with a low barrier of entry to those with less Python experience.

While collaborative data science is well documented in computational science literature, there are few mentions of truly collaborative citation analysis in library literature. Discourse remains centered on data literacy and data science support with a focus on retraining library workers, forming partnerships outside the library, and scrambling to assess "campus needs" as progress marches on (Burton et al. 2018; Oliver et al. 2019), though new work (Deardorff 2020) promotes collective efforts like Library Carpentries as a way to connect with and serve researchers. This study hopes to unify these strands by presenting a replicable path for citation analysis and data-driven collaborations among librarians.

Methods

Data acquisition

We used the subscription-based Clarivate Analytics Web of Science interface for initial data acquisition. We searched the Web of Science Core Collection database for all Earth Science articles and reviews published from 2010 to 2019 at each of the four authors' institutions. Using the advanced search interface, the publication searches for each institution were completed according to the parameters outlined in Table 1.

We downloaded search results as a tab-delimited file using the "Export Records to File" option, including both the Full Record and Cited References. This option allows for a maximum of 500 exported records at a time. We downloaded the data in batches of 500 and combined each separate file into one file per institution. The

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Table 1: Search parameters for initial document search.

Database:	Web of Science Core Collection
Document Types:	Article AND Review
Research Area (SU):[a]	Geology OR Geochemistry Geophysics OR Crystallography OR Meteorology Atmospheric Sciences OR Mineralogy OR Mining Mineral Processing OR Oceanography OR Physical Geography OR Water Resources OR Paleontology OR Remote Sensing
Organization Enhanced (OG):[b]	Authors' institutions (for example, "University of Houston")
Timespan:	2010–2019
Language:	English

a See Web of Science Core Collection Research Areas: https://images.webofknowledge.com/images/help/WOS/hp_research_areas_easca.html

b The Organization Enhanced field resolves name variants of institutions to a preferred organization name. See: https://images.webofknowledge.com/images/help/WOS/hp_organizations_enhanced_index.html

full search protocol can be found in Appendix 1. All subsequent data collection and analyses used programmatic techniques such as the Python programming language in the Jupyter Notebook interface to clean, transform, and refine the cited reference data. All of the code and documentation are publicly available at the project GitHub repository.[3]

Sampling

Using the Python programming language in a Jupyter Notebook, we created a stratified random sample (n = 1,000) based on each institution's proportion of the total records (Table 2). With the Pandas sample command, we then created a random sample for each institution and combined these proportional samples into one file for an aggregated 1,000 count sample.[4] We note that this study did not take into account instances of self-citation, which is a limitation of citation analyses and the bibliometric field as a whole.

3 https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project

4 The sampling process can be recreated using the following notebook: https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/Citation_Project_Sample.ipynb

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Table 2: Proportional Sample creation.

Institution	Number of Records	Proportion of Total	Sample (n = 1000)
Boulder	7439	0.42	420
Berkeley	5002	0.28	280
UCLA	3862	0.22	220
Houston	1344	0.08	80

The data exported from Web of Science contained 68 fields, including cited references for each publication. The cited references were listed in a single field with each cited reference delimited by a semicolon. We cleaned data by dropping unneeded fields, splitting the individual cited references for each publication into single fields, and then transforming the data from "wide format" to "long format" (i.e., one cited reference per row in a dedicated column rather than one publication per row with all of its cited references in separate columns). These steps were accomplished using the Pandas Python library and are documented in the project repository on GitHub.[5]

API usage

Cited references provided by the Web of Science are unstructured data, and to analyze citations to a given journal, the data needs to either be parsed or supplemented with outside metadata. Individual elements of the unparsed cited reference data (title, date, authors, etc.) are not split into their own fields. Splitting is complicated by some fields being present in certain citations and absent in others, making parsing difficult. We elected to supplement the cited reference data with metadata from CrossRef.[6] Eighty-one percent of the cited references contained a digital object identifier (DOI). Using a regular expression, we split the DOI of each cited reference into its own field. Then taking the cited references' DOIs, we queried the CrossRef REST API[7] to add clean, supplementary metadata fields to each citation. To accomplish this, we developed a script that iterated over each cited reference sending a query to the CrossRef API containing its DOI. Each query was structured as a url in the format described in Figure 1. From the CrossRef API response, we mined the citation title, publication name, date, and ISSN fields and added these to our cited reference data.[8]

5 https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/Citation_Project_Sample.ipynb

6 https://www.crossref.org

7 https://github.com/CrossRef/rest-api-doc

8 See cells 11 and 12 in https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/Citation%20Data%20Clean%20and%20API.ipynb for more information.

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

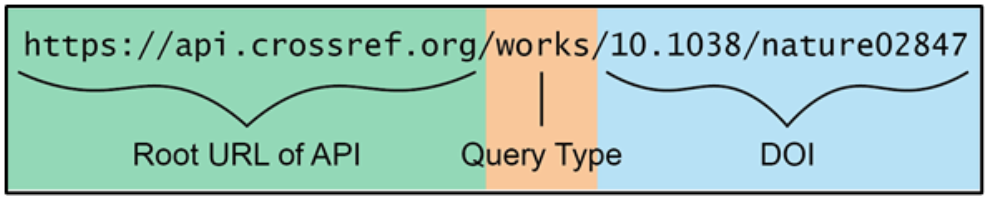

Figure 1: Structure of a CrossRef API Query.

80/20 rule calculation

Of the 81% of all cited references with DOIs, 99% had metadata available in CrossRef. We tested the 80/20 rule for these cited references by calculating the following new fields:

- 'counts': count of the number of times a publication was cited;
- 'cumsum': cumulative summation of the cited publications;
- 'titlenum': rank of each publication from most to least-often cited;
- 'titlepct': percentage of all citations attributed to each publication;
- 'citationpct': cumulative percentage of citations attributed to each publication.

We determined the percentage of publications cited by 80% of all of the cited references by counting the number of titles that cumulatively account for 80% of all citations, then calculating what percentage these titles represent among all of the titles. We then plotted the percent of total citations against the percent of total publications cited. We performed these procedures on the combined cited reference data from all four institutions as well on each individual institution to allow for inter-institutional comparisons. These methods can be reproduced in the notebook, "Citation_Data Analysis_All.ipynb".[9]

We explored the remaining 19% of citations that could not be matched with a DOI, but this was not the main focus of our analysis. A separate methodological explanation can be found in Appendix 3.

Additional querying through Jupyter Notebooks

We examined the journal titles most frequently published in by institution, the percentage of open access (OA) articles published in the same period, and the age of citations at the time of publication. Web of Science's Open Access (OA) designations are provided by Our Research, an organization that manages a knowledgebase of OA content and "makes it possible to discover and link to legal Gold or Bronze (free content at a publisher's website) and Green (self-archived in

9 https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/Citation_Data_Analysis_All.ipynb

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

a repository) OA versions" (Clarivate n.d.). We took the OA analysis a step further by considering what proportion of the cited references in our sample were OA. Cited references' OA statuses are not included in Web of Science's cited references data, but we could determine these articles' open access statuses by querying the Unpaywall.org REST API, also offered by Our Research.[10] The Unpaywall.org API accepts queries consisting of articles' DOIs and returns JSON-formatted information such as whether an article is OA, it's OA status (e.g., gold, green, etc), and more. These data were obtained for each of the 81% of cited references with DOI information and were sorted into two categories: cited references published OA and cited references behind a paywall. We then calculated the percentages of OA vs non-OA cited references for each institution over the 10-year study period.[11]

Results

We began by examining publishing output from the four institutions during our ten-year study period (2010-2019). Boulder affiliates published 7439 articles, the most of the four institutions. Berkeley published 5002, UCLA, 3862, and Houston, 1344. Despite the range of output, researchers at all four institutions showed similar journal preferences. The top ten journals show substantial overlap (Table 3). Boulder, Berkeley and UCLA researchers published in Geophysical Research Letters (GRL) most frequently. GRL also appeared in Houston's top ten list. The four titles highlighted in gray appeared in the top ten list for each institution. Houston's top title, Geophysics, published by the Society of Exploration Geophysicists, reflects the institution's focus on energy exploration and geophysics.

Citation Trends and the 80/20 rule

We then asked whether researchers in Earth Science rely on a smaller subset of literature than estimated by the 80/20 rule. Our 1,000-article sample yielded 55,580 citations (an average of 55.58 citations per article); 10,635 of those had no DOI and 11,280 lacked a standardized title. We analyzed the 44,300 citations with titles to evaluate our methods' reproducibility while minimizing time spent on data cleaning.

The 44,300 citations yielded a total of 2,715 distinct journals cited by our researchers. Table 4 shows the top five most frequently cited journals during the study period, *Journal of Geophysical Research*, *Geophysical Research Letters*, *Atmospheric Chemistry and Physics*, *Journal of Climate*, and *Science*.[12] Researchers at each individual campus showed a strong preference for journals published by the American Geophysical Union, with the Journal of Geophysical Research and its spin-off journals appearing on each institution's list.

10 https://unpaywall.org/products/api

11 This process is further documented in the following Jupyter notebook https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/open_access_analysis.ipynb

12 For a list of the top 250 titles, visit: https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/top_250.csv

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Table 3: Top 10 journals published in by campus taken from original publication data pre-sampling.

Rank	Boulder		Berkeley		UCLA		Houston	
	Title	Count	Title	Count	Title	Count	Title	Count
1	GEOPHYSICAL RESEARCH LETTERS	843	GEOPHYSICAL RESEARCH LETTERS	461	GEOPHYSICAL RESEARCH LETTERS	568	GEOPHYSICS	77
2	ATMOSPHERIC CHEMISTRY AND PHYSICS	643	ATMOSPHERIC CHEMISTRY AND PHYSICS	153	JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES	134	ATMOSPHERIC ENVIRONMENT	62
3	JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES	600	WATER RESOURCES RESEARCH	150	ATMOSPHERIC CHEMISTRY AND PHYSICS	123	ATMOSPHERIC CHEMISTRY AND PHYSICS	53
4	ATMOSPHERIC MEASUREMENT TECHNIQUES	204	EARTH AND PLANETARY SCIENCE LETTERS	133	JOURNAL OF CLIMATE	99	JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES	45
5	JOURNAL OF CLIMATE	190	JOURNAL OF GEOPHYSICAL RESEARCH-SOLID EARTH	132	CLIMATE DYNAMICS	94	EARTH AND PLANETARY SCIENCE LETTERS	30
6	BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY	139	GEOCHIMICA ET COSMOCHIMICA ACTA	128	EARTH AND PLANETARY SCIENCE LETTERS	88	GEOCHIMICA ET COSMOCHIMICA ACTA	28
7	WATER RESOURCES RESEARCH	116	JOURNAL OF GEOPHYSICAL RESEARCH-ATMOSPHERES	100	GEOCHIMICA ET COSMOCHIMICA ACTA	78	GEOPHYSICAL RESEARCH LETTERS	26
8	ATMOSPHERIC ENVIRONMENT	111	GEOPHYSICAL JOURNAL INTERNATIONAL	96	METEORITICS & PLANETARY SCIENCE	73	IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING	25
9	MONTHLY WEATHER REVIEW	105	ENVIRONMENTAL RESEARCH LETTERS	92	JOURNAL OF GEOPHYSICAL RESEARCH-PLANETS	65	IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING	25
10	EARTH AND PLANETARY SCIENCE LETTERS	95	ATMOSPHERIC ENVIRONMENT	72	JOURNAL OF GEOPHYSICAL RESEARCH-OCEANS	63	CRYSTAL GROWTH & DESIGN	25

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

There is some overlap between the journals that researchers publish in and those they cite. Researchers both read and publish frequently in the Journal of Geophysical Research (including its spin-offs), Geophysical Research Letters, Atmospheric Chemistry and Physics, and Earth and Planetary Sciences Letters. Due to their competitive acceptance rates, researchers are more likely to cite from than to appear in prominent journals like Science and Nature. Researchers within Earth Science disciplines also show a preference for society-published journals, the most heavily cited being titles published by the American Geophysical Union, European Geosciences Union, and to a lesser extent, the American Meteorological Society. Atmospheric Chemistry and Physics is the only gold open access journal in Table 4.

Table 4: Top 5 cited journals.

Rank	Aggregate	Boulder	Berkeley	UCLA	Houston
1	Journal of Geophysical Research	Journal of Geophysical Research	Journal of Geophysical Research	Geophysical Research Letters	Earth and Planetary Science Letters
2	Geophysical Research Letters	Geophysical Research Letters	Geophysical Research Letters	Journal of Geophysical Research	Geochimica et Cosmochimica Acta
3	Atmospheric Chemistry and Physics	Atmospheric Chemistry and Physics	Earth and Planetary Science Letters	Journal of Climate	Geophysics
4	Journal of Climate	Journal of Geophysical Research: Atmospheres	Science	Science	Meteoritics & Planetary Science
5	Science	Journal of Climate	Nature	Journal of Geophysical Research: Space Physics	Journal of Geophysical Research

The Journal of Geophysical Research (JGR) posed a quandary. From 1896-1977, it was a single journal. From 1977 onward, JGR began to split into various titles, but indexing systems often retained the original name of Journal of Geophysical Research. As a result, determining the top cited journals required sifting through instances of Journal of Geophysical Research, which was not an easy task. A closer examination of the cited references data determined that DOIs of articles published prior to 2013 in any part of the journal were attributed to *Journal of Geophysical Research* rather than to specific sections. Since 2013, article DOIs have been assigned by section, such *as Journal of Geophysical Research: Solid Earth*. This case is somewhat specific to a subset of Earth Science, but it raises

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

broader considerations for future studies. Many fields rely on serials with complicated histories of mergers and splits. Disciplinary knowledge can aid in identifying these titles, spotting discrepancies, and evaluating the significance of the results.

After determining top-cited journals, we then calculated the 80/20 rule for this varied set of institutions. The aggregated results (Table 5) from our sample indicate that only 7.88% of the journals were responsible for 80% of the citations. White (2019) demonstrated that 10% of journal titles were responsible for 80% of citations at Boulder for the years 2012-2016. For this study period (2010-2019), 9% of Boulder's cited journals were responsible for 80% of the represented citations, generally replicating White's results.

Table 5: Citation counts and 80/20 rule representation.

	Aggregate	**Boulder**	**Berkeley**	**UCLA**	**Houston**
Total Citations	44,300	19,562	12,979	10,575	1,184
Total number of journal titles	2,715	1,426	1,595	1,185	299
% of journals responsible for 80% of citations	7.88%	9.18%	16.99%	15.02%	34.11%

For Berkeley and UCLA, 17% and 15% of journals respectively were responsible for 80% of the citations. Houston had the smallest proportion of citations in our sample, but also demonstrated the most varied set of citations with 34% of journals responsible for 80% of the citations. Figure 2 visualizes the 80/20 rule for our sample as a whole and for each individual campus.

Age of Citations

Earth Scientists at our institutions tended to cite new research with a much greater frequency than older materials (Figure 3). The mean article age at the time of citing was 13 years old. A few citations of very old materials skewed the mean upwards. More telling, however, was that 25% of all articles were just four years old or less, 50% were nine years old or less, and 75% of all cited articles were 17 years old or less. This result is generally in concurrence with White (2019), which reported a mean citation age at time of citing of nine years. Our data negates the conventional wisdom among Earth Science Librarians that Earth Scientists use a lot of older materials. While many Earth Scientists, particularly Geologists, may rely on a lot of older materials for their work, it is clear that Earth Scientists on the whole employ recently published research to inform their work.

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

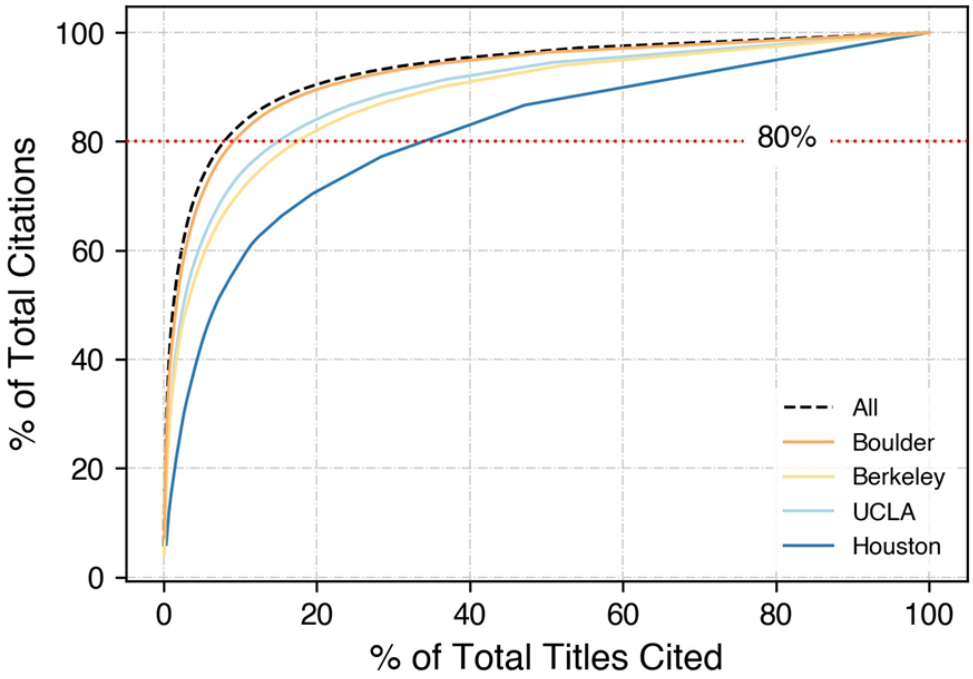

Figure 2: Representation of 80/20 rule: Total Citations vs. Total Titles Cited by Institution.

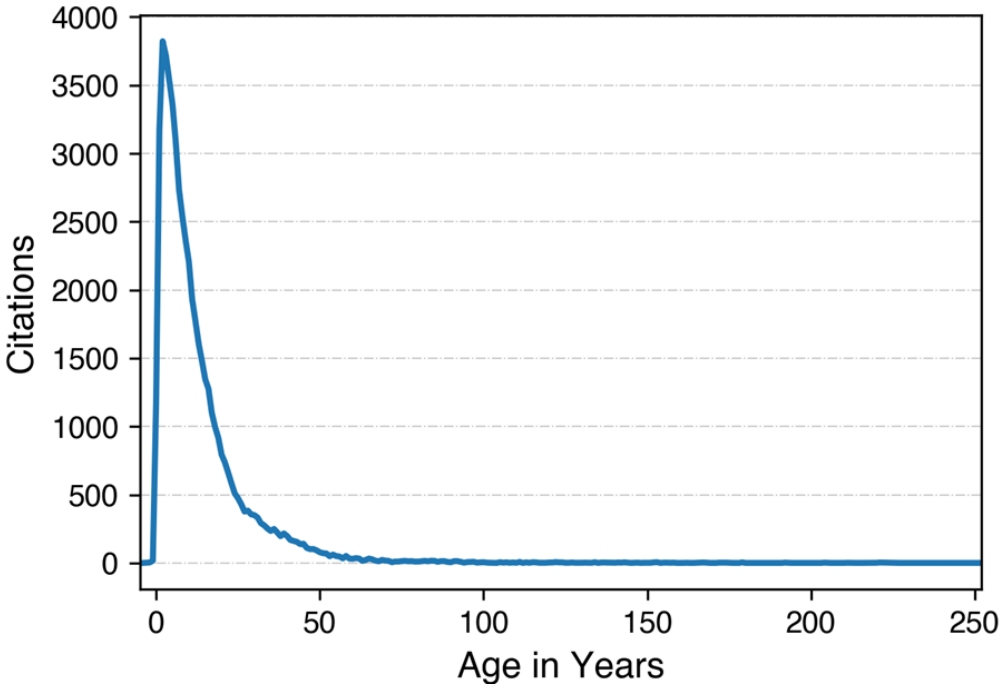

Figure 3: Age of Cited Reference at Time of Citation. The youngest item cited was -4 years old, while the oldest item cited was 276 years old (plot truncated at 250).

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Further analysis: Open Access

Determining how often researchers at our institutions publish open access was dependent upon the metadata assigned by Web of Science. By taking a count of each article's OA type by institution, we calculated the types of OA represented in the data set. Appendix 2 shows the variety of OA types represented, as well as overlap when articles utilize multiple types of OA. For the purposes of this analysis, we were less interested in the particular types of OA and more curious about the uptake of Open Access overall and the availability of a free version of the article in any form.

Figure 4 shows the percent of articles published as OA during the ten-year period for each institution. Boulder moved from a low of 36% to a peak of 70% in 2016; Berkeley from 26% to 55%, also peaking in 2016; UCLA from 28% to 64%, again peaking in 2016; and Houston, seemingly from no Open Access prior to 2016 to 29% in 2019.

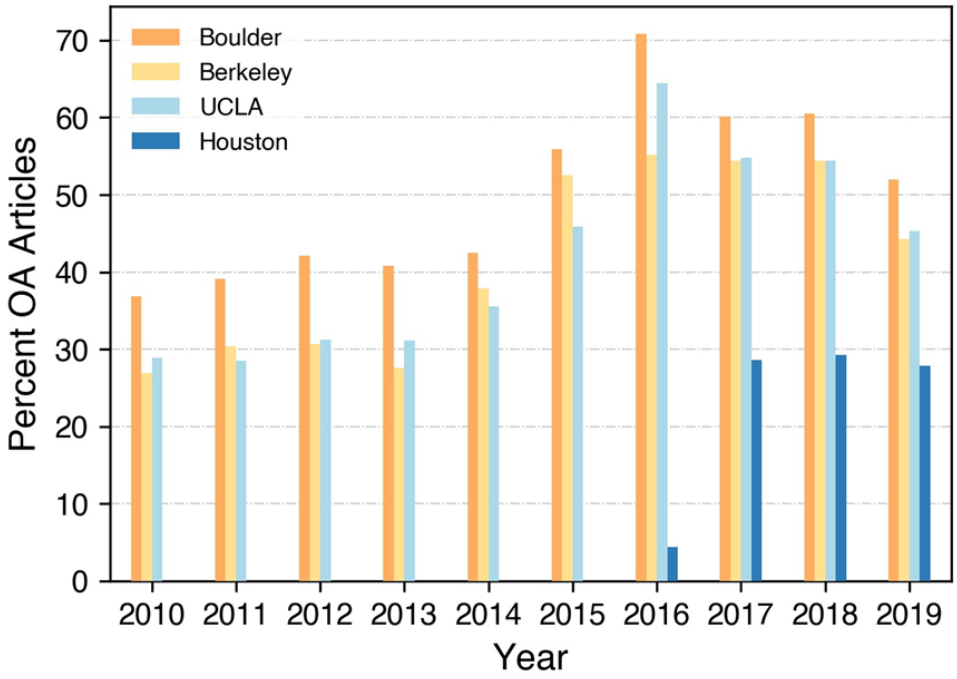

Figure 4: Open Access articles published by institution.

Figure 5 identifies the current Open Access Status of articles cited by Earth Science researchers by institution. Of the 55,580 cited references, 10,635 did not have a DOI and hence were not queryable. Unpaywall returned results for all but 407 of the 44,945 cited references with DOIs. The 10,635 cited references without DOIs and the 407 null results from Unpaywall were categorized as "Unknown," leaving 44,538 cited references whose OA statuses were determined. Proportions of citations to OA and non-OA articles suggest that Earth Science scholars rely

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

slightly more on paywalled articles. Boulder researchers cited OA articles the most frequently (42% of the time) while researchers at Houston cited OA the least frequently (23% of the time). Boulder was the only institution whose researchers cited OA articles more than paywalled articles. It appears that authors in the Earth Sciences seek source materials without preference or dependence on their OA status, as we find no clear trends in their citation habits.

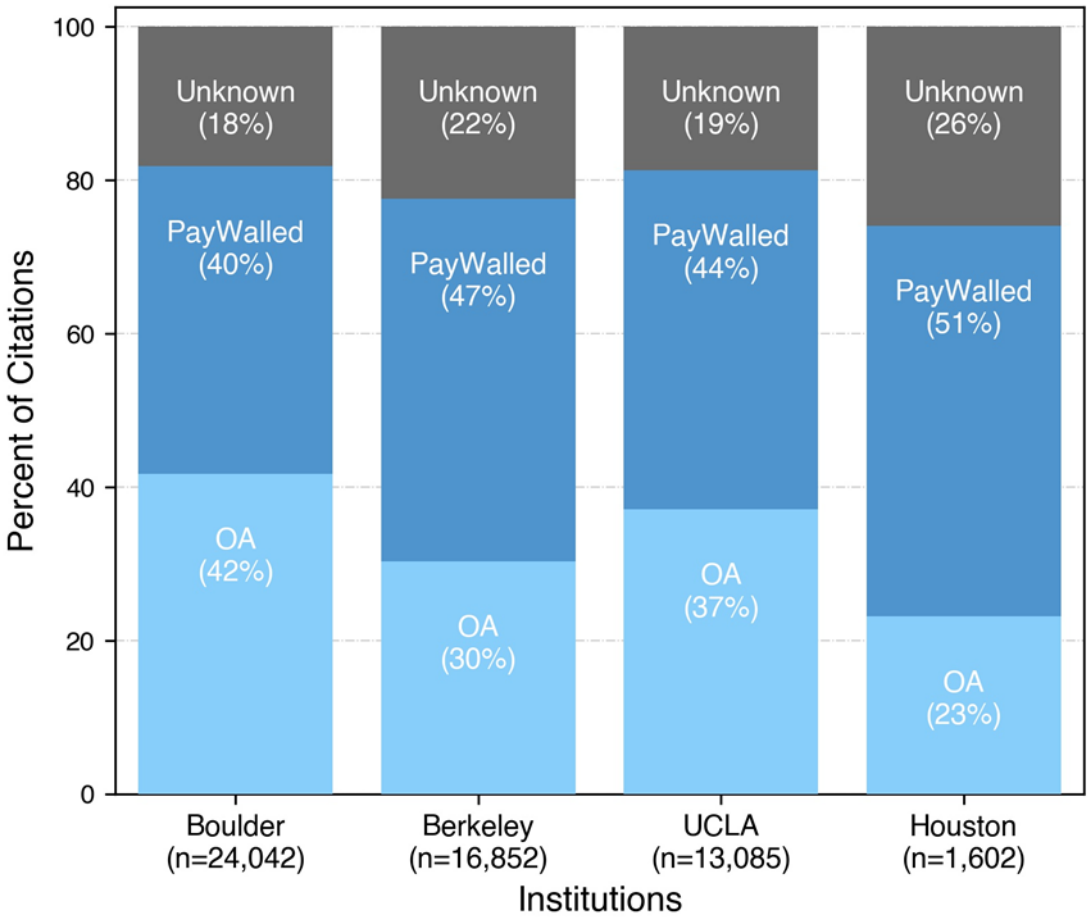

Figure 5: Open Access Status of Cited References by Institution.

We note however, the OA status of 21% of the cited references remain unknown. This uncertainty occurred because 20% of the total 55,580 citations lacked a DOI and another 1% that did have a DOI were absent from the Unpaywall database.

Perhaps similar proportions of the remaining 21% are OA and paywalled. The absence of DOIs among these materials could also indicate older works, gray literature, government documents, or other materials that may not receive a specific designation of being OA or not. Examining these non-DOI cited references as well as trends over time could illuminate authors' habits in citing OA materials.

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

Discussion

Our results suggest broader implications regarding the limitations of known ideas such as the 80/20 rule, the uptake of open access in Earth Science publishing, and the application of open science methods.

Limitations of the 80/20 rule

The 80/20 rule provides a useful tool to identify core collecting needs, but limitations are apparent. Individual institutions and departments within them specialize in different topics. Applied locally, the 80/20 rule varies according to campus research interests, and in the case of some institutions like Houston, results may differ due to the small representation in the sample, being a smaller department overall, or perhaps broader research interests that lead to less concentration on a small set of journals. Our results also suggest that larger sample size leads to a smaller portion of journal titles being responsible for 80% of the citations. With greater sample depth, a higher proportion of titles cited at a low frequency are present in the data. This phenomenon may partially explain why the smallest sub-sample (Houston) has the most varied title representation at the 80% margin (34%) while the largest sub-sample (Boulder) has the least varied titles (9%) at the same margin—as well as why the aggregate percentage of journals responsible for 80% is even less (7.8%).

Based on our analysis, a comprehensive Earth Science collection could include just 200 serial titles and sufficiently cover the needs of most researchers at most institutions.[13] At the opposite end, the citation data has an extremely long tail of items cited only once. Further analysis could investigate these items but is unlikely to yield sufficient insights for collecting purposes. For smaller institutions, and institutions where the 80/20 rule goes beyond 20%, a close examination of frequently cited works beyond the top 100 titles might inform distinct collecting that fulfills the needs of the institution's specific research focus.

Open Access in the Earth Sciences

Prior studies covered overlapping time periods between 2009 and 2015 and found anywhere from 30-50% of Earth Science articles to be available in different forms of OA (Piwowar et al. 2018; Archambault et al. 2014; Martín-Martín, Costas, Van Leeuwen, et al. 2018). We extend that time period through 2019 and found between 30-70% of articles in the study to be openly accessible depending on the institution. Our results tentatively suggest an upward trend in OA publishing among Earth Scientists over the past decade, in concurrence with Piwowar et al. (2018).

OA policies implemented in the mid-2010s may have influenced this trend. For example, the University of California passed an open access policy covering faculty

13 For a full list of our top 250 cited titles, visit: https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project/blob/master/top_250.csv

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

in 2013, and a subsequent Presidential policy in 2015. (Office of Scholarly Communication, University of California, n.d.). Similarly, the faculty of the University of Colorado, Boulder adopted an open access policy in 2015 encouraging submission of works to the university's institutional repository (University of Colorado Boulder, n.d.). Houston started discussions on supporting researchers publishing open access a few years ago, and is still in the process of making a formal policy. Authors who participate in the curriculum vitae review service can submit pre- or post-prints to be placed in the institutional repository for open access. This service began in 2018.

The slight downward trend following a 2016 peak in Figure 4 is worth noting. We speculate that the recent downturn may be related to OA "embargos" enacted by some publishers. It is also unclear in Clarivate's documentation how embargos are dealt with in designating articles' OA statuses in their data (Clarivate n.d.). Further work in this area would necessitate delving into individual publishers' OA policies and embargo practices over time. Mandates likely play a role in open access preferences as well. Earth Science researchers in the United States are more likely to be funded by NASA, DOE, and NSF than early OA supporters like NIH. Further study might consider the influence of project funders on OA publishing trends in this field.

While our analysis was effective in determining if an article is published open access, it is more challenging, and perhaps less relevant, to determine the open status of articles at the time of citation. Researchers have many means of accessing an article, including their own institutional subscriptions, collaborators and colleagues, institutional repositories, and various other sites across the Internet. It would be impossible to determine how or even if a researcher consulted a full article before citing it, but tracking support for Open Access at the journal level (as in the case of Atmospheric Chemistry and Physics, see Table 4) could signal changing publishing trends and preferences in a given discipline. For example, the combined monitoring of OA publishing and OA citation would help demonstrate interest in and use of diamond journals like the new family of titles from Tektonika and Seismica (https://twitter.com/WeAreTektonika, http://seismica.org) that require neither Article Processing Charges (APCs) nor subscription fees.

Methodology and Implications for Open Science

We began this project questioning whether we could apply White (2019)'s methods to programmatically assess the publication and citation practices of Earth Science researchers, while simultaneously offering an updated method that would be replicable on its own. We intended to use White (2019)'s methods with the addition of Jupyter Notebooks to track our analyses. There was some overlap between the data of the present study and White's original study data. White's 2019 study data used every citation from one academic department from the five-year period of 2012 through 2016. Our study used data from the same institution, but defined more broadly and encompassing multiple academic

JeSLIB 2021; 10(2): e1194
https://doi.org/10.7191/jeslib.2021.1194

departments. Further, the present study's sampling of the 10-year period of 2010 to 2019 means that while some data from the earlier study may have been included in the present study's data, there would have been minimum overlap.

The Python script used by White (2019) is available online through GitHub.[14] However, since Clarivate's Web of Science Database is not freely available and their API protocols have become more restrictive, exact reproducibility of methods was not possible. Whitaker (2017) and Clyburne-Sherin (2020) define replicability as using the same methods with different data, and also offer a generalizable solution with the addition of new methods. Hence, we confirm the results of White 2019, although we did not strictly replicate it. The two studies' results for Boulder are strikingly similar; 9% of Boulder's cited journals were responsible for 80% of the represented citations in the current study, and 10% in White 2019.

Our study introduced several different methods to further increase the replicability of our process. We used Jupyter Notebooks to create samples, to interact with the CrossRef API, and to create figures from our data. Our notebooks are available to reuse in a GitHub repository and act as a narrative of our workflow, following the guidelines of Rule et al., 2019. Results have also been exported from the Notebooks and saved in the GitHub repository to insure reproducibility. Going forward, researchers can replicate our methods by rerunning the Jupyter Notebooks with their own institutional data.

Conclusions

Jupyter Notebooks are a viable platform for disseminating replicable processes for citation analysis, making regular citation analysis less fraught and time consuming, ultimately leading to a deeper understanding of disciplinary research and publication habits. Notebooks are easy to share, but can be hard to collaborate on simultaneously because only one user can edit at a time. This may change as new platforms for collaboration emerge, but in the meanwhile, separating methods into distinct notebooks with distinct tasks, paying attention to version control via GitHub, and ironing out shared procedures ahead of time can be helpful in collaborative projects.

We were unable to construct a completely open methodology. WoS allowed us to isolate our discipline's research by institution and subject—a feature not provided by an open resource. At this point, an analysis using APIs and open indexes exclusively is still out of reach. However, the methods presented here are far more approachable and efficient than analysis by other means. Evolving products like OpenCitations may prove helpful to future work.

Our analysis also revealed that the average age of a cited paper is 13 years, indicating Earth Scientists at our institutions prefer recent work. Additionally, the top five cited journals were nearly identical in the case of Boulder, Berkeley, and UCLA. Houston was an outlier, likely due to sample size. In aggregate, 80% of the

14 https://github.com/outpw/WOKapiscripts

citations could be accounted for by only 7.88% of journals. When this number was broken down by institution, Boulder's citations proved to be the least diverse. Berkeley and UCLA had a very similar breakdown with slightly less than twenty percent of their citations comprising eighty percent of their journals. Again, Houston was the outlier, with well over twenty percent of their citations representing eighty percent of journals. In all categories, Berkeley and UCLA were the most similar.

Our choice of sampling method introduced some limitations. Using a non-stratified sampling technique would allow us to determine if similarities between Berkeley and UCLA were due to their similarly sized departments, or to factors such as increased collaboration or shared consortia. Furthermore, we originally chose the proportional sampling method to achieve more representative aggregate results, but the resulting small sample size for Houston may have skewed that campus' individual results. Future work might revise the methodology by using equal samples to achieve more accurate results for local analysis.

Future work could also delve more deeply into Open Access publishing and citation preferences, considering the type of Open Access most prevalent and the specific timeline of citation—in the case of embargoed articles, were they cited under subscription or open access. This has collection implications but can be hard to deduce programmatically. Regardless of the limitations described, we feel this methodology offers an approachable and reproducible process for librarians who are just beginning to explore data-driven research.

Acknowledgements

The authors contributed to this project in the following roles:
ST: Conceptualization, Data Curation, Methodology, Writing, Validation, Visualization, Project Management.
WT: Conceptualization, Investigation, Validation, Writing.
MW: Conceptualization, Writing, Validation.
PW: Conceptualization, Methodology, Software, Visualization, Data Curation, Writing.

Supplemental Content

Appendices
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2021.1194 under "Additional Files".

Data Availability

Data and code are available at the following GitHub repository and will be preserved on Zenodo upon publication: https://github.com/samteplitzky/Earth-Science-Citation-Replication-Project

References

Antelman, Kristin. 2004. "Do Open-Access Articles Have a Greater Research Impact?" *College & Research Libraries* 65(5): 372–382. https://doi.org/10.5860/crl.65.5.372

Archambault, Éric, Didier Amyot, Philippe Deschamps, Aurore Nicol, Françoise Provencher, Lise Rebout, and Guillaume Roberge. 2014. "Proportion of Open Access Papers Published in Peer-Reviewed Journals at the European and World Levels—1996–2013." *Copyright, Fair Use, Scholarly Communication, etc.* 8. https://digitalcommons.unl.edu/scholcom/8

Arendt, Julie, Bettina Peacemaker, and Hillary Miller. 2019. "Same Question, Different World: Replicating an Open Access Research Impact Study." *College & Research Libraries* 80(3): 303–318. https://doi.org/10.5860/crl.80.3.303

Borowski, Marcel, Johannes Zagermann, Clemens Klokmose, Harald Reiterer, and Roman Radle. 2020. "Exploring the Benefits and Barriers of Using Computational Notebooks for Collaborative Programming Assignments." In *SIGCSE*, 468–74. Portland, OR, USA. https://dl.acm.org/doi/pdf/10.1145/3328778.3366887

Burton, Matt, Liz Lyon, Chris Erdmann, and Bonnie Tijerina. 2018. "Shifting to Data Savvy: The Future of Data Science in Libraries." Project Report. Pittsburgh, PA: University of Pittsburgh. http://d-scholarship.pitt.edu/id/eprint/33891

Clarivate. n.d. "Open Access." Open Access. n.d. https://incites.help.clarivate.com/Content/open-access.htm

Clyburne-Sherin, April, and Seth Ariel Green. 2020. "T10: Open Source Tools: Train-the-Trainer Course." *Open Science Framework*. https://doi.org/10.17605/OSF.IO/X64GB

Deardorff, Ariel. 2020. "Why Do Biomedical Researchers Learn to Program? An Exploratory Investigation." *Journal of the Medical Library Association* 108(1). https://doi.org/10.5195/jmla.2020.819

deVries, Susann, Robert Kelly, and Paula M. Storm. 2010. "Moving beyond Citation Analysis: How Surveys and Interviews Enhance, Enrich, and Expand Your Research Findings." College & Research Libraries 71(5): 456–466. https://doi.org/10.5860/crl-45r1

Eckman, Charles. 1988. "Journal Review in an Environmental Design Library." *Collection Management* 10(1–2): 69–84. https://doi.org/10.1300/J105v10n01_07

Fleming, Thomas P., and Frederick G. Kilgour. 1964. "Moderately and Heavily Used Biomedical Journals." *Bulletin of the Medical Library Association* 52(1): 234–241. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC198101

Frohlich, Cliff, and Lynn Resler. 2001. "Analysis of Publications and Citations from a Geophysics Research Institute." *Journal of the American Society for Information Science and Technology* 52(9): 701–713. https://doi.org/10.1002/asi.1121

Helama, Samuli. 2012. "A Review of Citation Patterns in Doctoral Dissertations at the Department of Geology, University of Helsinki, Finland, since 1896." *Science & Technology Libraries* 31(2): 180–189. https://doi.org/10.1080/0194262X.2012.676870

Hoffmann, Kristin, and Lise Doucette. 2012. "A Review of Citation Analysis Methodologies for Collection Management." *College & Research Libraries* 73(4): 321–35. https://doi.org/10.5860/crl-254

King, Molly M., Carl T. Bergstrom, Shelley Correll, Jennifer Jacquet, and Jevin West. 2016. "Self-Citation and Gender." *Open Science Framework*. https://osf.io/de853

Martín-Martín, Alberto, Rodrigo Costas, Thed van Leeuwen, and Emilio Delgado López-Cózar. 2018. "Evidence of Open Access of Scientific Publications in Google Scholar: A Large-Scale Analysis." *Journal of Informetrics* 12(3): 819–841. https://doi.org/10.1016/j.joi.2018.06.012

Nabe, Jonathan, and Andrea Imre. 2008. "Dissertation Citations in Organismal Biology at Southern Illinois University at Carbondale: Implications for Collection Development." *Issues in Science and Technology Librarianship* 55(Fall). https://doi.org/10.5062/F46W980N

Nisonger, Thomas E. 2008. "The 80/20 Rule and Core Journals." *The Serials Librarian* 55(1–2): 62–84. https://doi.org/10.1080/03615260801970774

NSF—National Science Foundation. n.d. "Public Access - Special Report." https://www.nsf.gov/news/special_reports/public_access

Office of Scholarly Communication, University of California. n.d. "Participate in the UC open access policies." Accessed October 15, 2020. https://osc.universityofcalifornia.edu/for-authors/open-access-policy

Oliver, Jeffrey C., Christine Kollen, Benjamin Hickson, and Fernando Rios. 2019. "Data Science Support at the Academic Library." *Journal of Library Administration* 59(3): 241–257. https://doi.org/10.1080/01930826.2019.1583015

Perez, Fernando, and Brian E. Granger. 2015. "Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science." Grant Proposal funded by the Helmsley Trust, the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation. UC Berkeley and Cal Poly. http://archive.ipython.org/JupyterGrantNarrative-2015.pdf

Perkel, Jeffery M. 2018. "Why Jupyter Is Data Scientists' Computational Notebook of Choice." *Nature* 563(7729): 145–146. http://dx.doi.org/10.1038/d41586-018-07196-1

Piwowar, Heather, Jason Priem, Vincent Larivière, Juan Pablo Alperin, Lisa Matthias, Bree Norlander, Ashley Farley, Jevin West, and Stefanie Haustein. 2018. "The State of OA: A Large-Scale Analysis of the Prevalence and Impact of Open Access Articles." *PeerJ* 6(February): e4375. https://doi.org/10.7717/peerj.4375

Rule, Adam, Amanda Birmingham, Cristal Zuniga, Ilkay Altintas, Shih-Cheng Huang, Rob Knight, Niema Moshiri, et al. 2019. "Ten Simple Rules for Writing and Sharing Computational Analyses in Jupyter Notebooks." *PLOS Computational Biology* 15(7): e1007007. https://doi.org/10.1371/journal.pcbi.1007007

Sennyey, Pongracz, Gillian D Ellern, and Nancy Newsome. 2002. "Collection Development and a Long-Term Periodical Use Study: Methodology and Implications." *Serials Review* 28(1): 38–44. https://doi.org/10.1080/00987913.2002.10764705

Sterman, Leila, and Jason Clark. 2017. "Citations as Data: Harvesting the Scholarly Record of Your University to Enrich Institutional Knowledge and Support Research." *College & Research Libraries* 78(7): 952–963. https://doi.org/10.5860/crl.78.7.952

Stoudt, Sara, Valeri N. Vasquez, and Ciera C. Martinez. 2020. "Principles for Data Analysis Workflows." ArXiv:2007.08708 [Cs]. http://arxiv.org/abs/2007.08708

Trueswell, Richard. 1969. "Some Behavioral Patterns of Library Users: The 80/20 Rule." *Wilson Library Bulletin* 43(5): 458–461.

University of Colorado Boulder. n.d. "Campus Open Access Policy." Accessed October 15, 2020. https://www.colorado.edu/policies/campus-open-access-policy

Walcott, Rosalind. 1992. "Characteristics of Citations in Geoscience Doctoral Dissertations Accepted at United States Academic Institutions 1981-1985." *Science & Technology Libraries* 12(2): 5–16. https://doi.org/10.1300/J122v12n02_02

Whitaker, Kirstie. 2017. "Publishing a Reproducible Paper." *figshare*. https://doi.org/10.6084/m9.figshare.5440621.v2

White, Philip B. 2019. "Using Data Mining for Citation Analysis" *College & Research Libraries* 80(1): 76–93. https://doi.org/10.5860/crl.80.1.76

Zipp, Louise S. 1996. "Thesis and Dissertation Citations as Indicators of Faculty Research Use of University Library Journal Collections." *Library Resources & Technical Services* 40(4): 335–42. https://doi.org/10.5860/lrts.40n4.335