Scientometrics (2022) 127:4977–5003
https://doi.org/10.1007/s11192-022-04467-7

Impact factions: assessing the citation impact of different types of open access repositories

Jonathan Wheeler[1] · Ngoc-Minh Pham[2] · Kenning Arlitsch[3] · Justin D. Shanks[4]

Received: 2 February 2022 / Accepted: 30 June 2022 / Published online: 29 July 2022

Abstract

Institutional repositories (IR) maintained by research libraries play a central role in providing open access to taxpayer-funded research products. It is difficult to measure the extent to which IR contribute to new scholarship because publisher self-archiving policies typically require researchers to cite the "version of record" of a manuscript even when an IR copy is accessed to conduct the research. While some studies report an open access (OA) citation advantage resulting from the availability of self-archived or "green" OA manuscripts, few have sought to measure an OA citation effect of IR separately from disciplinary repositories, including arXiv and PubMed Central. In this study, the authors present a bibliometric analysis examining correlations between search engine performance of items in IR, OA availability from different types of repositories, and citations. The analysis uses a novel, open dataset of IR access and usage derived from five months of Google search engine results pages (SERP) data, which were aggregated by the Repository Analytics and Metrics Portal (RAMP) web service. Findings indicate that making OA copies of manuscripts available in self-archiving or "green" repositories results in a positive citation effect, although the disciplinary repositories within the sample significantly outperform the other types of OA services analyzed. Also evident is an increase in citations when a single manuscript is available in multiple OA sources.

Keywords Institutional repositories · Search engine performance · Bibliometrics · Open access citation advantage · Open access availability

✉ Jonathan Wheeler
jwheel01@unm.edu

[1] University of New Mexico, Albuquerque, NM, USA

[2] School of Information Science and Learning Technologies, University of Missouri, Columbia, MO, USA

[3] Central European University, Vienna, Austria

[4] Ingredients Consulting, Bozeman, MT, USA

Introduction

Institutional repositories (IR) are a common and critical component of the services provided by academic libraries. Established, conceptually, in the late 1990s as a hedge against rising serial costs by facilitating open access (OA) to research, the scope of IR services has grown to include the promotion and dissemination of scholarly works, datasets, administrative records, electronic theses and dissertations, and other content. As of May 2022, the OpenDOAR registry provided by Jisc lists 5862 institutional repositories around the globe that provide access to a broad range of content types across disciplines (Jisc, 2022).

Collectively, IR represent a significant global investment in OA infrastructure but their success in countering rising serials costs is debatable (Poynder, 2016). The return on investment for an IR at a single institution can be difficult to assess. The diversity of available platforms, capacity for customizing local configurations of open source platforms, and the absence of standardized impact metrics have led to a varied IR ecosystem across which it is difficult to define consistent benchmarks and standards of performance (Arlitsch & Grant, 2018). Additionally, differences among the most commonly used methods for gathering statistics can result in dramatically overcounting or undercounting statistics such as file downloads, page visits, and number of unique visitors within a given timeframe (OBrien et al., 2016). Bot activity in IR has been shown to comprise as high as 85% of all traffic (Greene, 2016) and it complicates the calculations of web metrics because methods used to filter bots from human web activity are not consistently applied across platforms. In the IR ecosystem, for example, a primary option is the default bot filtering capabilities of the IR platform, itself, which can be impacted by local configuration and depends on the capabilities and maintenance schedule of its developers. Alternatively, IR may participate in analytics projects such as IRUS-UK or IRUS-US, which apply standardized COUNTER bot filtering specifications (Lambert & Needham, 2019; Needham & Stone, 2012). Bot filtering heuristics must account for behaviors including the number of times a page is visited, duration of visits and the frequency of user behaviors including clicks, scrolling, etc. (Greene, 2017).

Particularly troubling, from an administrative and IR advocacy standpoint, is that current practices make it difficult to assess the contributions IR make to the use and citation of the scholarly works they contain. Compliance with commercial publisher policies for publishing within IR the submitted or accepted manuscripts of “version of record” articles typically requires citation of the paywalled, published copy.[1] This practice allows publishers of paywalled content to leverage open and often publicly-funded IR infrastructure for the benefit of their own impact factors (González-Betancor & Dorta-González, 2019). It also creates an ethical problem, as the taxpayers who fund research may be unable to access the paywalled products of that research. Finally, compliance with publisher self-archiving policies undermines the sustainability of IR because their positive effects may be hidden. If open access to scholarly literature is to remain a cornerstone of the IR value proposition, then it is necessary to recognize that compliance with publisher self-archiving policies is self-defeating and undermines the long-term sustainability of IR.

It is important to develop methods to quantify the value of IR as services that promote scholarship. Research suggests that IR contribute to an increase in citations for openly

[1] For example, see the Sherpa Romeo entry for the Journal of Academic Librarianship, https://v2.sherpa.ac.uk/id/publication/14175, which indicates that the accepted copies of articles uploaded to IR must link via DOI to the publisher version of the article (Jisc, nd.).

accessible scholarly content. This phenomenon is understood as a potential OA citation advantage. A sub-category of the OA citation advantage, referred to here as a green OA citation advantage, expresses the citation impact of self-archiving the submitted (pre-print) or accepted (post-print) manuscripts of published articles within IR and discipline-specific repositories such as arXiv[2] and PubMed Central.[3] We refer to these latter repositories throughout this article as "disciplinary repositories."

The research presented here differentiates between IR and a sample of disciplinary repositories, publisher-provided, and other types of OA services in a bibliometric analysis of correlations between citations and the search engine performance of items held by IR. Specifically, we focus on click activity, which we define as the number of clicks from search engine results pages that are received by URLs pointing to items within IR.

Literature review

An association between OA and increased citation is found in numerous studies. The size of reported citation advantages has narrowed over time as methods and definitions of open access have evolved since Lawrence's initial discussion of citation advantages for articles that were freely available online (Lawrence, 2001). Regardless of how broadly or narrowly "open access" is defined, robust cases for an OA citation advantage have been described by numerous studies (Abbasi et al., 2019; Antelman, 2004; Gargouri et al., 2010; McCabe & Snyder, 2014; Piwowar et al., 2018; Razumova & Kuznetsov, 2019; Xia et al., 2011). Ottaviani (2016) provides a succinct summary of positive OA citation advantage findings over time, suggesting a real but likely modest citation advantage for open access content. More recently, a systematic review of 134 studies showed that nearly half confirmed the existence of OACA while one quarter found that it did not exist (Langham-Putrow et al., 2021).

Findings in support of an OA citation advantage are robust across methodologies and disciplines. A range of methods and metrics have been used in studies affirming the existence of an OA citation advantage. Commonly used measures include average citations and citation counts. Abbasi et al. (2019) determined that the ratio of the average citations between OA and non-OA articles is 15.6:2.25. Archambault et al. (2016) present data on the average of relative citations for 3.3. million papers published from 2007 to 2009 and indexed in the Web of Science (WoS). These data show a decidedly large citation advantage for OA papers, despite a lag in availability of OA compared to paywalled papers among the papers included in their study. Arendt et al. (2019) used citation counts and observed similar patterns to those found by Antelman (2004): freely accessible articles receive more citation counts than paywalled. Other metrics used to measure a statistically significant OA citation advantage include: citation rate (Alkhawtani et al., 2020; Antelman, 2004; Bautista-Puig et al., 2020), and risk of not being cited (Eysenbach, 2006). Models used in the study of an OA citation advantage include logistic regression (Eysenbach, 2006; Gargouri et al., 2010), stepwise backwards linear regression (Eysenbach, 2006), and negative binomial regression (Fraser et al., 2020).

An OA citation advantage is also found across disciplines. Antelman (2004) looked at articles in four disciplines (i.e., philosophy, political science, electrical and electronic

[2] https://arxiv.org/.

[3] https://www.ncbi.nlm.nih.gov/pmc/.

engineering, and mathematics) with varying degrees of growth and prevalence of OA to see whether OA articles have greater impact as measured by citations. Results showed that freely available articles have a greater research impact, but the observed citation advantage was not evenly distributed across disciplines. The discipline with the largest growth in OA availability in Antelman's study was mathematics, but the discipline with the greatest impact of OA on citation rates was political science (Antelman, 2004). Similarly, Arendt et al. (2019), Björk et al. (2010), and Hajjem et al. (2006) report a general citation advantage for open access papers that is sensitive to disciplinary context but nonetheless consistently demonstrated across disciplines.

Despite research suggesting a measurable OA citation advantage, there is also evidence against it. For example, a recent analysis of citations among OA and paywalled journal articles found that there is no general OA citation advantage and that article access status accounts for little of the variability in the number of citations an article receives (Basson, 2019). Instead, confounding factors including the reputation of the journal, the language the journal is published in, and the first author's home institution can have a stronger effect on the number of citations an article receives (Basson, 2019; Dorta-González et al., 2020).

A detailed overview of such confounding factors and their effect on prior studies is provided by Craig et al. (2007). In particular, two alternative explanations for a perceived OA citation advantage deserve special note. The first, selection bias, posits that open access papers receive more citations because the cost and effort involved in self-archiving or paying article processing fees results in authors choosing to make only their best work open access. Craig et al. (2007) provide a thorough discussion of this issue and the effect on OA citation advantage studies when analyses do not account for selection bias. A second alternative explanation for the increased citations received by OA papers is the early view postulate (Craig et al., 2007), which asserts that papers made available through preprint servers like arXiv receive more citations because they are able to have more immediate exposure and impact when compared with the delay in scholarly publishing that can affect paywalled papers. While both of these factors merit special consideration, a 2010 study by Gargouri confirmed an OA citation advantage that was independent of confounding factors. Instead, results suggested that a measurable, independent OA citation advantage exists, with highly cited articles receiving the most benefit (Gargouri et al., 2010).

Further confounding a clear understanding of any OA citation advantage are the varying and broad definitions of OA (Beatty, 2019). Arendt et al. (2019) followed Antelman's (2004) operational definition of "open" as free availability, as indexed by Google. Antelman uses "free access" or "freely available" rather than "open access" to describe articles in the dataset. She counted articles as free if they were accessible directly by clicking on the link from Google's results page, or if the results led to a page that contained article metadata and a link, such as one labeled "PDF" or "Full text," that led to a free access copy. This is a broader definition of "open access" than the one provided by the Berlin Declaration (Max Planck Institute, 2003), and it can inflate the citation impact of open access as a philosophical principle. It's possible to overstate the importance of distinguishing between types of OA, especially as researchers may be satisfied with obtaining free access to content, regardless of whether the content has been made intentionally green or gold OA. However, in addition to free access, models of OA including green and gold further provide for reuse and preservation beyond the limitations of traditional copyright. These additional features of OA relate to the value of IR as green OA repositories, highlighting the importance of established definitions of "open access" to scholarly communications research.

Studies of the scholarly impact of open access provided by IR may be less sensitive to questions of definition, however, as IR are commonly understood to provide green OA. Importantly, a citation advantage attributable specifically to green OA has been reported by Young and Brandes (2020), and Archambault et al., (2014, 2016). Archambault identified the use of IR and immediate green OA in order to avoid embargo delays and showed there is a significant citation advantage to green OA publishing (Archambault et al., 2014, 2016). Specifically, papers in general science and technology, historical studies, and visual and performing arts all receive, on average, twice as many citations as the overall population of papers (Archambault et al., 2014). Despite explicating the role of IR in green OA citation advantage, Archambault et al. (2016) did not detail or specifically measure the effect of IR on an OA citation advantage.

Research purpose

Our dataset consists only of items available through some form of OA. We cannot therefore test for an overall OA citation advantage. However, as summarized in the literature review there is an argument to be made for an OA citation advantage, and a green OA advantage in particular. It is thus worth assessing the comparative citation rates of items held within IR and other types of OA repositories and services.

The specific role of IR in contributing to a potential OA citation advantage is an underdeveloped research area. This is possibly due to the noted difficulty of measuring actual research use of IR content. An exploration of the presence of IR content in 28 social tools using a webometric approach highlights this issue. The findings show that most IR have no strong presence in those tools, which included social media services like Facebook and Twitter, but also professional social networking tools like ResearchGate, LinkedIn, and Zenodo (Aguillo, 2020). Aguillo's research indicates that articles tend not be cited using the "URL of the IR that offers information about institutional authorship" (Aguillo, 2020). Comparing the citation rates of items held by IR with those of other types of OA services provides a means of assessing the value of IR as research repositories independently of the other types of OA providers.

In addition to separating IR from other types of OA services our research demonstrates the analytic potential of a novel dataset of IR use and performance from the Repository Analytics and Metrics Portal (RAMP)[4] (Arlitsch & Wheeler, 2020; OBrien et al., 2017). RAMP is distinct from other IR metrics services because it harvests data from Google Search Console (Google, 2020) for each participating IR's pages and hosted content files returned from searches on Google properties (e.g., web search, image search, Google Scholar). For example, RAMP captures data when users access PDF files directly from a Google Scholar search result page rather than navigating to the item from within the IR. RAMP data can be considered supplementary to page-tagging analytics platforms like Google Analytics or log-based metrics. The data are of research interest because RAMP applies a consistent set of metrics for all repository platforms, solving the problem of potentially large variation between usage statistics reported by server logs versus those reported by page tagging services (OBrien et al., 2016, 2017). For a comparison of RAMP

[4] https://rampanalytics.org.

characteristics with those provided by page tagging and server logging tools, please see (OBrien et al., 2017).

The RAMP service has been available for free registration to repository managers since January 2017.[5] As of May 2022, RAMP includes over 70 repositories from around the globe, representing all of the major IR software platforms of both general-purpose IR and disciplinary repositories. RAMP administrators encourage use of the data for analysis of IR search engine optimization and benchmarking, as well as research about IR content and access. To support such research, we published a subset of RAMP data, consisting of augmented search performance data about the content of 35 participating repositories between January 1 and May 31, 2019 (Wheeler & Arlitsch, 2020; Wheeler et al., 2020). This publicly available dataset was used for the analyses reported in this study. A larger dataset of 2017–2021 RAMP data has subsequently been published as annual subsets in the Dryad data repository (Wheeler & Arlitsch, 2021a, 2021b, 2021c, 2021d, 2021e).

RAMP data can be merged with complementary datasets to explore new methods for assessing the role and contribution of IR to an "open scholarly record," which is here defined as the global set of academic, research, and other scholarly products that are openly accessible. The research described in this article combines publicly available RAMP data with citation information from Crossref (*Crossref REST API*, n.d.) and OA availability data from Unpaywall (*Unpaywall REST API*, 2020). The study assesses correlations between citations and the search engine performance of individual items hosted in IR by investigating how differences in search engine performance of IR items correspond to differences in OA availability and citations. Specifically, we address three research questions:

1. Do items that receive higher numbers of clicks in search engine result pages have higher rates of citations than other items?
2. How does the availability of items from different types of OA repositories affect citation counts?
3. How does the availability of multiple copies of an item within different types of OA repositories affect citation counts?

We note that overall search engine performance can be measured via multiple statistics. Our study is specifically focused on the number of clicks on URLs in search engine result pages (SERP) that point to content files of items hosted by IR.

Methods

In addition to reviewing findings related to IR and the OA citation advantage, a further objective of the literature review was to identify sources of bibliometric data that can be merged with RAMP data to enable cross-comparison between bibliometric and search engine performance data. Bibliometric studies use various citation databases: Scopus (Abbasi et al., 2019); Web of Science (WoS) (Antelman, 2004; Arendt et al., 2019; Beatty, 2019); Journal Metrics (Abbasi et al., 2019); SCImago (Abbasi et al., 2019); DOAJ database (Bautista-Puig et al., 2020); Open Access Directory (OAD) (Bautista-Puig et al., 2020); analytics tool like the 1science OAIndx (Archambault et al., 2016); and discipline

[5] Repository managers may register at https://rampanalytics.org.

specific websites (Alkhawtani et al., 2020). Additionally, previous OA citation studies extracted citation indicators from multiple sources. For example, the research population in Abbasi et al. (2019) comprised LIS journals and articles in LIS hybrid journals in Scopus. The data related to citation indicators (number of received citations, two year's impact, Citescore [IPP], and H-index) were extracted from Scopus, Journal Metrics, and SCImago.

Analyses of an OA citation advantage can make use of generic, non-discipline specialized databases such as Scopus, WoS, or discipline-oriented databases. For example, Antelman (2004) and Basson (2019) each used data from generic, non-discipline specialized databases. Antelman used mean citation rates (as recorded in the ISI Web of Science database) of OA articles compared with those of non-OA articles for a sample population of journal articles in four disciplines. Basson conducted an analysis with all articles and reviews published from 2005 to 2014 and indexed in the Clarivate Analytics Web of Science (WoS). The OA citation advantage study from Alkhawtani et al. (2020) used data (including citation numbers) extracted from the discipline-specific database European Radiology website.

The literature review thus identified several commonly used bibliometric data sources, most notably Web of Science, Scopus, and Google Scholar. Although these data sources occur frequently in the OA citation advantage literature, we did not use data from Web of Science or Scopus because they do not provide an open, publicly accessible application programming interface (API) (Piwowar et al., 2018). Google Scholar, although easily accessible, does not provide an API, and Google discourages scraping data from Scholar SERP.

Alternatively, Crossref is a citation data source referenced in the literature, which does provide an open, public API. Crossref is a robust source of data for bibliometric research (Fraser et al., 2020; Hendricks et al., 2020; Piwowar et al., 2018), which we chose as the source of citation data because it enabled integration of automated data harvesting into the data aggregation workflow, along with the improved transparency provided by public data access.

In addition to citation data, a correlation between citations and the number of OA copies of any given article has been described within the OA citation advantage literature (Xia et al., 2011). Unpaywall is a source of information about OA availability, and a public API is available. Along with data about where OA copies of articles are hosted, Unpaywall further designates one copy as the "best OA" option, identifies which copy of an article is made available by each host (pre-print, post-print, etc.) and classifies hosts as either a "publisher" (for gold OA) or a "repository" (for green OA). See Fraser et al. (2020) for a description of use of Unpaywall data for OA availability research.

Data collection and aggregation

We performed our analysis using the publicly available subset of RAMP data (Wheeler et al., 2020) collected from 35 participating repositories between January 1 and May 31, 2019. Our objective was to compare the number of clicks on URLs pointing to content files of manuscripts published in IR with citations received by the corresponding articles to determine if availability from IR and other types of OA repositories is correlated with clicks in SERP. In addition to detailed documentation provided with the dataset, information about how data are harvested and processed for indexing in RAMP is available (Wheeler & Arlitsch, 2020). As noted in the published documentation, two sets of data are harvested each day for each IR participating in RAMP. Only one of these daily harvests includes data at the granularity of individual URLs. These are the data used in the analyses.

Data are processed prior to indexing in RAMP to identify URLs that point specifically to non-HTML content files (i.e., PDF, CSV, etc.). Tracking SERP performance of content files rather than item HTML pages is useful because clicks on content files may indicate a higher degree of user interest than a view of the HTML page, which generally contains only the abstract and metadata about the item (OBrien et al., 2017). For the purposes of this study, an "item" is defined as any work, together with its metadata and corresponding content files, that has its own HTML landing page in an IR.

RAMP data were further processed prior to analysis to account for two factors. First, an item within an IR may include multiple content files, each of which has its own URL. Second, any content file URL may occur in SERP multiple times during the period of study. To arrive at a set of unique items that received clicks on content files, URLs of items in RAMP that received at least one click within this timeframe were processed to infer the HTML URL of the file's parent item within the host IR. The resulting HTML URLs were further deduplicated to account for variations resulting from a repository's use of both insecure (HTTP) and secure (HTTPS) pages, as well as platform updates or similar changes that can result in a single item being referenced via multiple URLs.

We performed additional processes prior to the bibliometric analysis. RAMP data consist only of SERP information about items held by IR and the data are harvested from Google Search Console. The data do not by themselves include descriptive metadata or information that can be directly merged with Crossref or Unpaywall data. Additional pre-processing steps were:

1. Item level metadata were harvested from the HTML pages of parent items within IR that contained content files which received clicks from SERP during the period of study
2. DOIs, where available, were extracted from item level metadata along with the date of each item's publication within its host IR
3. Citation and OA availability for extracted DOIs data were retrieved from Crossref and Unpaywall
4. RAMP data for each unique item in the subset were aggregated to determine the total number of clicks received on URLs pointing to content files belonging to the item. Item level RAMP data were then merged with DOI and date of IR publication metadata using the parent item's HTML URL as a shared identifier
5. The combined RAMP data and item level metadata were merged with citation and OA availability data, using each item's DOI as a shared, unique identifier.

We note that the process of deduplicating and aggregating RAMP data to a set of unique parent item URLs introduces limitations. The structure of the RAMP data required significant pre-processing to arrive at the set of unique HTML pages of items containing content files. This included reverse engineering the parent item's HTML page for all content files in the dataset. The process varies by repository platform, and features of some repository platforms may result in multiple items being aggregated as a single item. For example, some repositories have a "recently added" feature that creates a single HTML "item" feed, which can include ephemeral links to content files from multiple different parent items. Using the aggregation process developed for this study, clicks on any of these temporary links would be aggregated under the feed URL rather than the actual parent item's HTML URL. Content files appearing in SERP may also be paginated in a way that results in inflating the number of clicks received by content file URLs associated with individual items, though we note the effect of potential outliers are addressed in the ANOVA analysis.

The data were also filtered prior to metadata harvesting and analysis to include only items with content files that received one or more clicks. As a result, this may bias the findings relative to the correlation between counts of OA copies of items hosted by IR and citations, since a large number of items that received zero clicks were excluded. This approach was taken for two reasons. First, to reduce the burden on IR servers of extensive web scraping, which can affect their performance and result in connections being forcibly closed. Second, items without clicks were not "used," as defined here, within the study period. Without data about the specific search queries that resulted in IR content appearing in SERP, the inclusion of items with zero clicks might have introduced unknown factors relating to why items were or weren't clicked into an analysis of whether recorded use correlates with citations.

Further information about RAMP data is available from the published dataset (Wheeler et al., 2020). Additional description of the data aggregation process is available with the analysis code from GitHub, https://github.com/imls-measuring-up/ramp_citation_analysis.

Regarding our analysis of OA availability, we note that Unpaywall data include a classification of OA hosts by type, either "publisher" or "repository." This classification does not distinguish institutional repositories from other types of green OA providers, most notably established and highly visible disciplinary repositories like PubMed Central and arXiv, the disciplinary repositories primarily represented within our sample. To enable a more granular analysis of how different types of green OA hosts may impact citations, all hosts occurring within the harvested Unpaywall data were manually classified as either "institutional," "discipline," "publisher," or "other" subtypes. This allowed the research team to include counts of OA copies by OA host subtype in the analyzed dataset. We note that the Unpaywall data do not include counts of copies of items hosted by ResearchGate, Academia.edu, Mendeley, or other research social media services. This may result from copies of items in the dataset not being available from those services at the time Unpaywall data were harvested. Alternatively, Unpaywall may not have indexed those services during the period of data collection. Our analysis therefore does not extend to these types of academic social media services.

Additionally, not all of the repositories included in the study had been indexed by Unpaywall at the time data were harvested from that service. As result, there are cases in which Unpaywall reported zero OA copies of items even though every item in the dataset has at least the one OA copy available from the RAMP participating IR whose data were included in the study. To address this, the counts of total OA copies and copies hosted by IR were increased by 1 for any item for which the hosting RAMP IR was not included in the Unpaywall data for the corresponding DOI.

The manual classification of OA hosts into repository subtypes introduces another potential limitation of our findings. All OA hosts classified as "publisher" were assigned the same subtype ("publisher"). Subtypes for OA hosts classified as "repository" within the Unpaywall data were identified based on the author's knowledge of academic institutional repositories, and through web searches and OpenDOAR. Repositories were given an "other" subtype in cases where a determination could not be made. This potentially biased citation analysis findings relative to "institutional" and "other" subtypes. Since the "other" subtype is something of a catch-all, results for this subtype should be treated with caution.

Prior to analysis, in addition to dropping observations with any null values we also dropped observations for items with IR publication dates later than 2016. Although this considerably reduced the size of the dataset, we removed these observations to reduce bias

Table 1 Count of items by year of upload to RAMP IR host

Year uploaded to IR	Count	Proportion
2004	2	0.01
2005	35	0.26
2006	1162	8.63
2007	228	1.69
2008	310	2.30
2009	326	2.42
2010	730	5.42
2011	575	4.27
2012	913	6.78
2013	1373	10.20
2014	1919	14.26
2015	2355	17.50
2016	3529	26.22

against items that had been published for less than two years prior to data collection. The two-year limit was based on methods of previous studies (Gargouri et al., 2010) which allow for a leveling off of citation rates within STEM fields after two years, and also to account for the delay between the publication of an article and its citation in other studies. Since the RAMP data were collected in early 2019, removing from the dataset any items that were published after December 31, 2016 means that all items within the dataset were published at least two years prior to our data collection. Table 1 provides a count of items in the final dataset broken down by the year in which they were uploaded to the RAMP IR whose copy appeared in a Google SERP during the period of study.

We also dropped observations for items that had an IR publication date that was earlier than the creation date of the Crossref DOI, or for which the IR publication date was more than a year after the creation date of the Crossref DOI. This was done to exclude items that may have been published and received citations before being added to Crossref, and conversely to exclude items that may been added to an IR after citation rates had peaked and leveled out. Dropping observations for the above reasons resulted in the removal of all data for five of the RAMP IR represented in the dataset. The final sample size of the analytic dataset, in terms of the number of RAMP IR represented, is therefore 30 IR instead of 35. We note that throughout the analyses below, the count of IR hosts of OA copies includes the RAMP IR whose copy appeared in a Google SERP, as well as the other non-RAMP IR that host additional OA copies of manuscripts represented within the dataset.

Figure 1 provides an overview of the data aggregation process, including the size of the dataset after each step. The final dataset used for the analysis is included in the GitHub repository referenced above.

Analytic methods

The unit of analysis in our study is a manuscript of a scholarly research article hosted by an IR, as represented by the Crossref DOI associated with its published version. The dataset also includes other types of content, such as electronic theses and dissertations, for which the DOI references the copy of the item hosted by the parent IR.

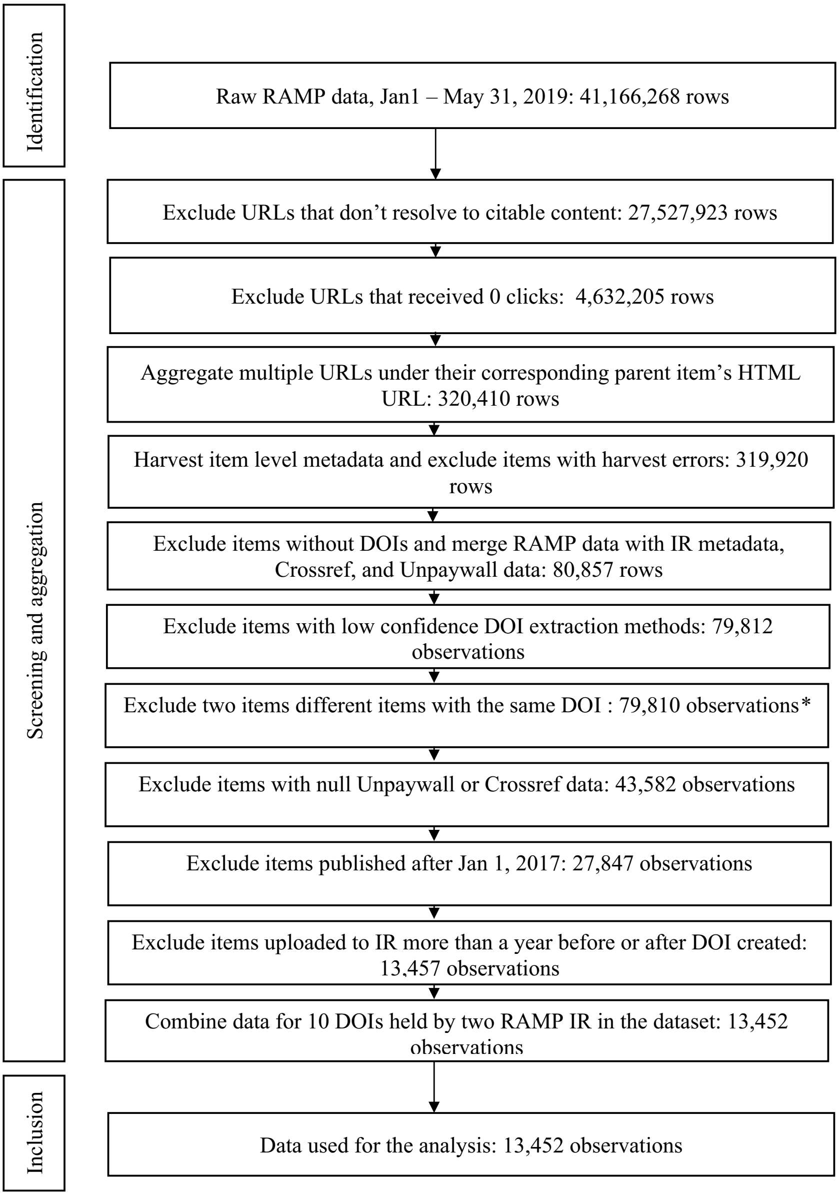
Identification
Raw RAMP data, Jan1 – May 31, 2019: 41,166,268 rows
Screening and aggregation
Exclude URLs that don’t resolve to citable content: 27,527,923 rows
Exclude URLs that received 0 clicks: 4,632,205 rows
Aggregate multiple URLs under their corresponding parent item’s HTML URL: 320,410 rows
Harvest item level metadata and exclude items with harvest errors: 319,920 rows
Exclude items without DOIs and merge RAMP data with IR metadata, Crossref, and Unpaywall data: 80,857 rows
Exclude items with low confidence DOI extraction methods: 79,812 observations
Exclude two items different items with the same DOI : 79,810 observations*
Exclude items with null Unpaywall or Crossref data: 43,582 observations
Exclude items published after Jan 1, 2017: 27,847 observations
Exclude items uploaded to IR more than a year before or after DOI created: 13,457 observations
Combine data for 10 DOIs held by two RAMP IR in the dataset: 13,452 observations
Inclusion
Data used for the analysis: 13,452 observations

Fig. 1 Flow chart showing the overview of data collection and selection process. Provided count represent the number of rows or observations remaining following each described step. *Remaining observation at this step are included in the published dataset. Additional data processing steps are performed in the published R script

We employed linear regression models as the analysis of variance (ANOVA) and the analysis of covariance (ANCOVA) to analyze correlations between citations, the total number of clicks on item URLs from SERP, and availability of OA copies of items from different types of repositories. The analysis was conducted in two steps. First, to find out whether the differences in the means of the citation rate between items grouped according to categories based on total clicks are statistically significant, we used linear regression analysis as one-way ANOVA. Second, multiple linear models were used as ANCOVA to investigate associations between citation rates per year (the response variable) and the availability of OA copies of items from different types of OA repositories (the categorical predictors) while considering the effects of clicks from SERP (the covariate). Results from both analyses were similar, so we report the results of the ANOVA models below.

The presence of outliers, normality, and variance homogeneity (ANOVA and ANCOVA assumptions) and homogeneity of covariate regression coefficients (ANCOVA-specific assumption) were tested before finalizing the models. Since clicks from SERP and citations are not evenly distributed across items, we cannot assume the residuals are normally distributed in each sub-group in any model. Therefore, we first examined the presence of outliers which may affect the interpretation of our models. Any observations with standardized residuals greater than 3 in absolute value were removed from the models as possible outliers. Next, we tested the normality assumption of the residuals, which were violated in all the models. The uneven distribution of citations within each sub-group and between the sub-groups defined for our analysis may explain this. Similarly, the variance homogeneity assumption is also violated in all the models developed for our analysis except for the ANOVA model. However, ANOVA and ANCOVA are robust against violations of normality and variance homogeneity assumptions. Therefore, the results from our ANOVA and ANCOVA analyses are valid but should be treated with caution.

The assumption of homogeneity of covariate regression coefficients was tested for all the ANCOVA models. This ANCOVA assumption was met for the ANCOVA models in which institutional repositories, disciplinary repositories, and publisher provided OA ($p > 0.05$) are predictors and was not met for the ANCOVA models in which total open access availability and the availability from other OA services are predictors ($p < 0.001$ and $p < 0.05$ respectively). The significant interaction effect between clicks and the number of other OA copies and total number of OA copies of a manuscript again indicate that the results of these models should be treated with caution.

Finally, to determine if the results in the regression models are biased because of the aforementioned violations, we used heteroscedasticity-consistent (HC) standard errors to obtain unbiased standard errors of ordinary least squares coefficients. The findings of the linear regression models with robust standard errors show consistent results with our linear regression models which removed outliers.

The predictor variable in the linear regression as the one-way ANOVA model is click categories. Items in our sample were categorized into two groups depending on the number of clicks from SERP received by each item: items with clicks equal to or below the median of 3 clicks, and those with clicks above the median (4 and more). We used median rather than mean to define categories since the mean of clicks is biased by possible outliers (see Table 6).

The dependent variable of the multiple linear regression models as the ANCOVA model is the average count of citations of a manuscript per year. The independent variables were categorized into groups to measure (1) whether the availability of items from different types of OA repositories affects citation counts, and (2) whether the availability of multiple copies of an item within different types of OA repositories affects citation counts. To serve the former purpose, independent variables corresponding to repository types were categorized into two groups based on whether zero or more OA copies of an item are available from each type of repository. To serve the latter purpose, when any type of repository had at least one OA copy of a given item, counts of available copies were categorized into two

Table 2 Open access availability by host type ($N=13{,}452$)

OA host type	Frequency	Percentage of observations
Items with OA availability	13,452	100.00
Items hosted by one or more IR	13,452	100.00
Items also hosted by disciplinary repositories	3496	25.99
Items also hosted by publisher OA repositories	3999	29.73
Items also hosted by other types of OA repositories	3755	27.91

groups based on the median count. Similar to click data, the median rather than the mean of the number of copies was used since the mean may be affected by possible outliers in the number of copies hosted by different types of repositories.

In all our models, the adjusted R squared values were consistently low. This means our models can only account for small percentages in the variability of the outcome variable. According to Grace-Martin (2012), such models are still useful in detecting if there is a small but reliable relationship between the predictor(s) and the outcome variable. The low adjusted R values also suggest that there are other variables that would be much better predictors of citation effects. Future studies could look for those variables.

The analyses were completed using R and RStudio (R Core Team, 2020; RStudio Team, 2020). Tables were generated using the Flextable (Gohel, 2021) and Stargazer (Hlavac, 2018) packages.

Results

Descriptive statistics are provided to illustrate the context of the data. Following date filtering as described above and removal of observations with null values, our sample size was 13,452 items. Because of the nature of RAMP data as search engine performance metrics for items hosted by open access IR, all the items in the dataset are available as OA in some form and all items are hosted by at least one of the 30 IR represented in the RAMP dataset. Among the other kinds of OA repositories into which we categorized items, 3496 (26%) of the items in the dataset have additional OA copies hosted by disciplinary repositories, 3999 (30%) have additional OA copies hosted by publishers, and 3755 (28%) have additional OA copies hosted by other types of repositories. Details of OA host type frequency are reported in Table 2.

Besides the small percentage of items in our sample which have additional OA copies hosted by disciplinary repositories, publisher provided, and other types of OA hosts, of special note is the small number of disciplinary repositories represented. Table 3 provides a breakdown of copies of items in the sample available from disciplinary repositories. Since some manuscripts in the sample are available from multiple disciplinary repositories, the table total is greater than that provided in Table 2 above for disciplinary repositories. Crucially, we note that the majority of these additional copies are hosted by only four repositories: PubMed Central, PubMed Central Europe, and two instances of the arXiv repository; arXiv.org and the Cornell University arXiv mirror.

For further detail about the repository representation within our dataset, supplementary data tables provide counts of the analyzed sample of items available from each of the manually categorized OA host types included in the analysis. Counts for each repository

Table 3 Distribution of items across disciplinary repositories

Repository	Count
PubMed Central	2424
PubMed Central—Europe PMC	2375
arXiv.org	571
Cornell University—arXiv	504
Econstor—Econstor	15
Ludwig Maximilian University of Munich—Munich Personal RePEc Archive	9
OSF Preprints—LawArXiv	9
Indiana University—Digital Library Of The Commons Repository	7
BePress Biostats	6
bioRxiv	5
GESIS â€ " Leibniz Institute for the Social Sciences—Social Science Open Access Repository	4
University of Minnesota, USA—AgEcon Search	4
University of Pittsburgh—PhilSci-Archive	4
PhilPapers Foundation—PhilPapers	3
Animal Studies Repository	1
Cold Spring Harbor Laboratory—bioRxiv	1
Modern Language Association / Columbia University—Humanities Commons CORE	1
OSF Preprints—SocArXiv	1
Wellbeing Studies Repository	1

Note Some items as reported in this table are available from multiple disciplinary repositories, so the total count is greater than the 3496 deduplicated count reported in Table 2.

are further broken down by the year in which the corresponding item was uploaded to the RAMP participating IR whose copy received at least click from a SERP for a Google property during the period of study. Although this date may be different from the date at which copies were uploaded to other IR, disciplinary repositories, publisher provided OA, or other services, the IR upload date is reported in each of the supplementary tables because the IR upload date was used in the analysis to determine the average annual citations for each DOI.

Tables are provided as CSV spreadsheets in the following order:

- Online Resource 1 includes the repository/year breakdown of items across IR
- Online Resource 2 includes the same information for items available from disciplinary repositories
- Online Resource 3 provides this information for items available from publisher provided OA
- Online Resource 4 provides this information for items available from other types of OA services and providers.

Please note that a single item may be available from multiple OA providers, so the totals in the tables may exceed the deduplicated totals reported in Table 2 and Table 5, below.

Items that received no clicks during the period of study were removed from the sample as part of the data aggregation process. Reasons for this are discussed above, but as

Table 4 Citation mean differences across click groups

Click group	N	Mean citations	SD	Median citations	Min citations	Max citations
Median and below (or 1–3 clicks)	7422	4.35	14.07	1.9	0	805
Above median	6030	4.95	15.76	2.0	0	637

Table 5 Citation means by OA host type

Host	Category	N	Mean	SD	Median	Min	Max
All OA hosts	Median and below or 1–2 copies	8571	3.38	8.09	1.55	0	289
	Above median or 3 or more copies	4881	6.79	22.04	2.83	0	805
Institutional repositories	Median and below or 1 copy	10,963	4.26	14.36	1.78	0	805
	Above median or more than 1 copy	2489	6.19	16.78	2.78	0	437
Disciplinary repositories	0 copies available	9956	3.42	8.95	1.50	0	437
	1 or more copies available	3496	8.03	24.61	3.60	0	805
Publisher OA	0 copies available	9453	4.10	14.53	1.71	0	805
	1 or more copies available	3999	5.85	15.52	2.57	0	437
Other OA	0 copies available	9697	4.70	14.31	2.00	0	805
	1 or more copies available	3755	4.42	16.18	1.71	0	637

a result, every item in our sample received at least 1 click within the 5-month period of study. The median number of clicks across all items throughout the period of study was 3. Details of citation mean descriptive statistics based on item groupings by total clicks are reported in Table 4. The categorization of items by clicks received from SERP in relation to the median number of clicks received from SERP is also used as the predictor in the ANOVA analysis and as the covariate in the ANCOVA analysis. When comparing the means of citations between two groups we observe in the descriptive statistics that items with total clicks above the median have a higher citation mean than those for which the number of clicks from SERP is below or equal to the median.

To measure the relationship between citation rates and types of OA, we categorized the independent variables into groups. We categorized groups differently due to the differences in overall frequency of items hosted by different types of OA repositories. When analyzing items based on the total number of OA copies and the count of items hosted by IR, items were categorized into groups relative to the median number of items hosted by services within these categories. Of 13,452 items in our sample, 8571 (64%) have a total of 1–2 OA copies available, and 4881 (36%) have 3 or more OA copies available. Of 13,452 items hosted in institutional repositories in our sample, 10,963 (81%) have 1 copy hosted in IR and 2489 (19%) have copies available from 2 or more IR. Recall that OA availability information is harvested from Unpaywall, and the count of copies of items available from IR is not limited to RAMP participating IR but instead includes any IR indexed by Unpaywall at the time of data collection.

Because of the lower overall number of items in our dataset with additional copies available from disciplinary repositories, publisher provided OA, and other types of OA providers, items for corresponding analyses were categorized by whether there

Table 6 Average annual citation rates by click groups

Click group	Estimate	Standard error	*t* value	Pr(>|*t*|)
(Intercept)	3.700	0.065	56.875	0.0000***
Above median	0.246	0.097	2.525	0.0116*

Residual standard error: 5.589 on 13,352 degrees of freedom

Multiple R-squared: 0.0004772, Adjusted R-squared: 0.0004024

F-statistic: 6.375 on 13,352 and 1 DF, p-value: 0.0116

*Signif. codes: 0 < = '***' < 0.001 < '**' < 0.01 < '*' < 0.05 < '.' < 0.1 < '' < 1*

were zero copies available from these types of OA providers or whether there were one or more copies. Descriptive citation statistics based on OA availability of items in our dataset are provided in Table 5.

When examining differences in citation means across the different types of OA repositories and services, the overall trend is that an increase in the number of copies in different types of OA repositories corresponds positively with the citation mean. For example, as shown in Table 5, items that have a total number of OA copies higher than the median of 2 copies have a citation mean ($M = 6.79$, $SD = 22.04$) that is almost twice as high as the citation mean for items for which the total number of OA copies falls below the median (or 1 to 2 copies) ($M = 3.38$, $SD = 8.09$). There is one exception to the overall trend, which is other types of OA repositories that we were unable to definitively categorize as institutional, disciplinary, or publisher provided OA. When items with copies hosted in other types of repositories are split into two groups, a "zero" copy group and a "1 or more" copy group, the citation mean for the latter group ($M = 4.42$, $SD = 16.18$) is lower than that for the former group ($M = 4.70$, $SD = 14.31$).

The linear and multiple linear regression models provide further evidence of the trends illustrated by the descriptive statistics. Clicks were consistently related to higher annual citation rates when used as the main predictor in all the linear regression models as the one-way ANOVA models. Clicks were likewise related to higher annual citation rates when used as the covariate in the multiple linear regression models as the ANCOVA models. For example, taking the dataset as a whole, the results from the linear regression test show that there is a significantly higher citation rate of 0.25 citations per year for a manuscript in the group that has clicks above the mean ($t = 2.5$, $p < 0.05$). Details of the linear regression model results are presented in Table 6.

Multiple linear regressions were conducted to examine the citation effects of the number of copies of an item available from different types of repositories. When controlling for the effects of clicks, we see that the total number of OA copies has a positive correlation with citations. To be more specific, the change from the number of total OA copies from median or below (1–2 copies) to a total number above the median (3 or more copies) is correlated with a higher annual mean citation rate of 2.29 citations per year ($t = 22.611$, $p < 0.001$). That is, the mean of the annual citation rate of items for which the number of OA copies is greater than the median number of OA copies ($M = 4.97$, $SE = 0.10$) was 85% higher than the mean of those with a total number of OA copies equal to the median or below ($M = 2.69$, $SE = 0.08$). Results for citation correlations across types of OA repositories are provided in Table 7 and described in more detail below. ANCOVA tests showed similar results.

Table 7 Citation impact of additional OA copies of items held by repository type

	Dependent variable				
	Per-year citation rate means				
	(1)	(2)	(3)	(4)	(5)
Intercept	2.688 (0.078) *t* = 34.400***	3.396 (0.070) *t* = 48.341***	2.589 (0.075) *t* = 34.577***	3.168 (0.075) *t* = 42.011***	3.793 (0.072) *t* = 52.582***
Clicks above median	0.682 (0.098) *t* = 6.984***	0.350 (0.097) *t* = 3.596***	0.900 (0.098) *t* = 9.146***	0.468 (0.098) *t* = 4.768***	0.238 (0.097) *t* = 2.452**
Total OA copies above median	2.286 (0.101) *t* = 22.611***				
Count IR copies above median		1.418 (0.125) *t* = 11.352***			
Disciplinary repository OA available			3.271 (0.112) *t* = 29.236***		
Publisher OA available				1.471 (0.107) *t* = 13.782***	
Other OA services available					− 0.319 (0.108) *t* = − 2.959***
Observations	13,356	13,355	13,361	13,355	13,354

Table 7 (continued)

	Dependent variable				
	Per-year citation rate means				
	(1)	(2)	(3)	(4)	(5)
R^2	0.037	0.010	0.061	0.015	0.001
Adjusted R^2	0.037	0.010	0.061	0.014	0.001
Residual Std. Error	5.514 (df = 13,353)	5.576 (df = 13,352)	5.520 (df = 13,358)	5.564 (df = 13,352)	5.587 (df = 13,351)
F Statistic	259.741*** (df = 2; 13,353)	67.835*** (df = 2; 13,352)	431.382*** (df = 2; 13,358)	98.398*** (df = 2; 13,352)	7.568*** (df = 2; 13,351)

*Note:**p**p***$p < 0.01$

Citation correlation for institutional repositories

A change in the number of copies of an item hosted by IR from less than or equal to the median of 1 copy to a count of IR availability above the median (2 or more copies) is correlated with a higher annual mean citation rate of 1.42 citations per year ($t = 11.352$, $p < 0.001$). The mean of the annual citation rate of items with a count of IR hosted OA copies above the median ($M = 4.81$, $SE = 0.13$) was 42% higher than that of the mean annual citations of those with a total IR OA availability equal to or below the median ($M = 3.40$, $SE = 0.07$).

Citation correlation for disciplinary repositories

A change from zero availability of additional copies of an item from disciplinary repositories to having at least 1 additional copy available is correlated with a higher average citation rate of 3.27 citations per year ($t = 29.236$, $p < 0.001$). This means that the average annual citation rate of items for which at least one other copy is available from disciplinary repositories ($M = 5.86$, $SE = 0.11$) was more than twice as high as the mean of those for which no additional copies are hosted by disciplinary repositories DR ($M = 2.59$, $SE = 0.08$).

Citation correlation for publisher OA repositories

For items with additional copies available through publisher OA, the change from zero copies to having at least 1 additional copy hosted by publisher OA is correlated with a higher average citation rate of 1.47 citations per year ($t = 13.782$, $p < 0.001$). This means that the average annual citation rate of items with at least one copy available via publisher OA ($M = 4.64$, $SE = 0.11$) was 46% higher than the mean of the those for which additional, publisher provided OA is unavailable ($M = 3.17$, $SE = 0.08$).

Citation correlation for other OA services

Contrary to the aforementioned repository types, the change from zero additional copies of an item being available from other types of OA services to having at least 1 additional copy hosted by those services is correlated with a lower rate of − 0.319 average annual citations between the two groups ($t = - 2.959$, $p < 0.01$).

Given the uneven distribution of items within our sample across disciplinary repositories, the reported impact on citations of OA availability from these repositories should be treated with caution. This uneven distribution may not only bias the citation effect of disciplinary repositories themselves, but also the disciplines they represent. Though the disciplinary repositories included as OA hosts within our sample were randomly selected, most of the data correspond to items available from health-related repositories (2439 items, 68.63%), followed by STEM related repositories (1052 items, 29.6%). Data from other types of repositories account for a small fraction of our disciplinary data (63 items,

Table 8 Citation mean differences by discipline

	Dependent variable Per-year citation rates of items available from disciplinary repositories, by discipline
Intercept (STEM disciplines)	7.136
	(0.327)
	$t = 21.837$***
Medical and Health Sciences	0.917
	(0.394)
	$t = 2.329$**
Others	– 4.024
	(0.605)
	$t = -6.652$***

Note *p**p***$p < 0.01$

1.77%).[6] This may result from disciplinary differences in the perceived value of OA, the availability of OA repositories, mandates, and norms around sharing pre and post-prints. It may also indicate that our disciplinary sample is not representative and therefore, the results may be biased and cannot be generalized for all disciplinary repositories.

A second ANOVA analysis of citation rates across disciplines covered by the disciplinary repositories in our sample further suggests a difference in citation practices across disciplines. Results are provided in Table 8. Within our sample, health related repositories show an average annual rate of 0.917 ($t = 2.329$, $p < 0.01$) citations more than STEM related repositories, with a rate of 7.136 annual average citations ($t = 21.837$, $p < 0.01$). Other disciplines represented within our sample of disciplinary repositories have an annual average citation rate that is – 4.024 citations lower than the STEM repositories in our sample ($t = -6.652$, $p < 0.01$). The degree to which disciplinary citation effects are impacted by the prevalence of health science repositories in our sample, or whether they result from other factors, is an area of further study.

Discussion

Our findings are suggestive of an overall OA citation advantage, since it is primarily through open access that multiple copies of an article can be made available to begin with. We reiterate that our analyses do not test for the existence of an overall OA citation advantage, as every item within our dataset is openly accessible from at least one IR. However, the results of our analysis with regard to citations of items held by IR, disciplinary repositories, and publisher repositories indicate that open availability of multiple copies of research articles from different kinds of OA providers is correlated with increased

[6] Note that these counts include 58 items for which copies were available from multiple disciplinary repositories covering different disciplines, so the count of 3,554 items here is greater than the deduplicated count of 3,496 items given in Tables 2 and 5.

citations. By classifying green OA hosts into subcategories of institutional repository, disciplinary repository, and "other," the current analysis suggests that any actual green OA citation advantage, and indeed any OA citation advantage in general, is largely correlated with the self-archiving of manuscripts in specialized disciplinary repositories.

Referring to Table 2, we recall that every item in the dataset is available from at least one IR, whereas the proportions of items with additional copies available from disciplinary repositories and publisher provided OA is much lower, at 26% and 30%, respectively. As noted below, this is not an indication of the underutilization of disciplinary repositories and publisher OA, but rather a probable bias toward IR content resulting from the nature of RAMP data as an IR metrics service. But even with significantly lower representation within our dataset, the citation effects of both disciplinary repositories and publisher provided OA were higher than the citation effect of IR. The difference between IR and disciplinary repositories is especially marked, with the citation effect of disciplinary repositories being more than twice as much as that of IR.

We further note that the citation effect within our sample is not only driven predominantly by disciplinary repositories but also that the vast majority of items in our sample with copies in disciplinary repositories are available from variants of PubMed Central and arXiv. Along with a higher citation effect, this suggests stronger overall search engine performance of and interest in the content available from these repositories. Importantly, our sampling method was not inherently biased toward these specific disciplinary repositories. Although our data collection was initiated by aggregating Google Search Console data for 35 RAMP participating IR, the sample is otherwise random with regard to which items have additional copies hosted by other, non-RAMP IR or other types of OA services. Further research is warranted in order to more fully explore the uneven distribution of disciplinary repositories within our sample, including consideration of the effects of funder OA mandates in the case of repositories like PubMed Central as well as cultures of open science and the corresponding adoption of preprint servers like arXiv.

The dominance of PubMed Central and arXiv within our sample of disciplinary repositories raises questions about sustainability within the disciplinary repository ecosystem that also merit further research. While the RAMP service and related research are oriented toward IR metrics in the interest of the sustainability of these platforms and services, it is understood—as underscored by our findings—that disciplinary repositories play a vital role in the dissemination of open access research products. However, the different missions, funding sources, organizational affiliations and other variations among disciplinary repositories make it difficult to generalize bibliometric findings, even across a more diverse sample than ours. The heterogeneous nature of disciplinary repositories, in contrast with general purpose services like IR, suggest that the development of more weighted measures of impact accounting for the different characteristics of disciplinary repositories is needed.

Similarly, an implication of our findings as they relate to IR is the limitation of citations as measures of scholarly impact and the importance to the IR value proposition of identifying and measuring indicators of broader, community impacts. IR often contain many items that are non-scholarly, or at least are not peer reviewed. Items of this type may be used in support of policy development or community-based initiatives—worthwhile uses of IR content that do not necessarily lead to citations. By contrast, disciplinary repositories provide a focused and convenient access point for primarily scholarly literature specific to certain fields and have been shown, when analyzed independently, to drive citations (Fraser et al., 2020).

It is worth noting the positive correlation between clicks and citations, and this makes a case for search engine optimization (SEO) in repositories. Without doubt, the largest single factor driving discoverability and use of IR content is a presence of that content in search engine indices. Proper implementation of SEO techniques, as well as monitoring and managing search engine performance using tools like Google Search Console, continues to be crucial in helping to make repository content discoverable. While we do not make a causal connection to the effect that accessing OA items as demonstrated by clicks in SERP leads to citations, the correlation suggests that IR are fulfilling their valued role of providing open access to research. Conceivably, a download of a manuscript as recorded by click activity leads to a citation of that manuscript, or alternatively, a highly cited item receives clicks because the corresponding research is a focus of popular news stories or other heightened interest. Also, OA content available from IR may often be used in teaching as an alternative to using copies that are paywalled or subject to prohibitive license restrictions. In either case, IR make high-interest content available to the general public, researchers, and educators without subscriptions. Here, the nature of IR as general-purpose repositories provides a net benefit, since not every discipline is served by an established disciplinary repository.

The question of whether and how often researchers use IR copies but then cite the publisher's "version of record" remains difficult to assess. However, the modest but significant correlation between citations and the count of OA copies held by IR suggests that at least some use of IR content is research oriented. Further, we note that the slightly higher correlation between publisher provided OA and citations may indicate that increasing uptake of OA models among publishers could result in less use of IR content for scholarly research over time. This smaller correlation should not be understood as an argument against the value of IR, but rather that outside of supporting research for peer reviewed scholarship the ongoing value of IR may more directly lie in providing access to unique collections including electronic theses and dissertations, administrative records, and various types of institutional grey literature. The value of electronic theses and dissertations as potential drivers of IR use has been previously reported in (Arlitsch et al., 2020).

Nonetheless, even if the citation effect of IR is outpaced by other forms of OA, it remains true that providing green open access to scholarly research is a foundation of the IR value proposition and a key part of IR service models. We advocate for the citation of IR copies of research articles when they are used, in addition to citing the publisher's copy, as an ethical way to support and recognize the contribution of IR to research and the open scholarly record.

Limitations

The findings reported here are based on an analysis of a dataset that is limited in scope in several ways. First, the subset of RAMP data used for the analysis was limited to 35 IR, of which the data for five repositories was dropped for reasons described in the methods section. Although the repositories represent a variety of institutional sizes, levels of research activity, and organizational structures (including consortia, technical institutes, and large research universities), the set may not be representative of the variety of repositories in the US or internationally. Further, the set is not culturally representative and consists primarily of institutions from the US, Canada, Europe, and Australia. The majority of the content hosted by the 35 repositories is in English.

A second limitation inherent in the data is RAMP's use of Google Search Console as the sole data source. While this enables baseline comparison of consistent IR search performance metrics across repository platforms, information about access and use of

IR content through other means besides searches on Google properties is not available. For example, use of IR hosted content in MOOCs, or shared via social media or academic social media services including ResearchGate and Academia.edu, would not be counted within RAMP data. However, as recently reported by Macgregor (2019), the vast majority of IR traffic is driven by Google properties.

The frequency of groups in Table 5 reveals that the majority of the manuscripts in our sample have only 1 copy hosted by institutional repositories (10,963, 81%). Disciplinary repositories, publisher repositories, and other repositories are less fully represented within our sample. This does not imply that these types of repositories are underutilized but is a further limitation of the scope of RAMP as a service for IR to measure their search engine performance. We note however that RAMP is available without cost to any repository that wishes to register, and that it may be used equally by disciplinary repositories, publisher OA repositories, and other types of OA services. Participation in RAMP from a larger variety of repository types and OA services can broaden the dataset and strengthen further research into correlations between search engine performance and citation effects.

Differences in citation activity across disciplines may also have impacted our analysis, as more recently published items within the RAMP dataset may not have reached their peak rate of citation. For example, any items in the RAMP data that are related to research in the STEM fields and which were published before 2014 or 2015 may well have leveled off in terms of overall citation rates. In this case, they will have received the majority of citations ahead of the 2017 cutoff date used in our analysis. However, items published in non-STEM fields as early as 2010 or 2011 may not have reached their peak citation rate or received a majority of their citations before January, 2017. Breaking down the analysis by discipline was not possible using available metadata, but we note that any bias resulting from this limitation would logically understate rather than overstate the impact of IR availability on citations received by items across the overall RAMP dataset.

Conclusions

This study suggests that a citation effect is most correlated with specialized disciplinary repositories, although the availability of manuscripts in IR and publisher OA repositories also has a positive, but smaller effect on the number of times the manuscript is cited. The results regarding availability of manuscripts in IR and the positive correlation with citations should be encouraging to proponents of IR. However, citations are only one measure of impact and may not provide the most effective means of assessing the value of IR. Further investigation of how IR content is used to support economic development, local and regional policymaking, and the publication and preservation of unique collections is warranted through the use of altmetrics and analysis of citations within grey literature. As noted above, the public availability of electronic theses and dissertations via IR is a potentially significant contribution to open scholarship and is a subject worthy of more detailed study.

One objective of the research reported here has been to identify publicly available data that could be merged with RAMP data in order to enable the definition and analysis of new metrics describing the contribution of IR to the open scholarly record. The bibliometric analysis described in this article included the following external data sources:

- Descriptive metadata for items that received one or more clicks in Google SERP during the five-month period of study. Although only dates of IR publication and DOI metadata were used for the analysis, additional metadata about authors, titles, keywords, and abstracts may be useful for further analysis of search engine performance relative to metadata, and variation in metadata use across repositories and platforms.
- Crossref bibliographic and citation metadata.
- Unpaywall data about OA availability of individual items.

All Crossref and Unpaywall data necessary to reproduce or replicate the reported analysis have been aggregated into the published dataset. Complete or more current data are available from the public APIs of both services.

We encourage our colleagues to make use of the RAMP datasets we have published. Six datasets of IR use and performance data, spanning 2017–2021 are freely available on the Dryad data repository. Institutional repositories continue to be a significant endeavor of many research institutions and we believe the data produced by RAMP will contribute to our understanding of their value through measurement of their performance and analysis of their content.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-022-04467-7.

Author contributions All authors contributed to the study conception, design, and drafting of the manuscript. Material preparation, data collection and analysis were performed by JW and N-MP. All authors read and approved the final manuscript.

Funding The research leading to these results received funding from the Institute of Museum and Library Services under grant agreements LG-06-14-0090 (2014-2017), LG-72-18-0719 (2018–2020), and RE-70-17-0094-17 (2017–2019). Authors Wheeler and Arlitsch administer the RAMP service described in the article. The service is provided free of charge to interested parties and no salary or other funding is received from RAMP by either author.

Declarations

Conflict of interest Authors Pham and Shanks have no competing interests to declare that are relevant to this article.

References

Abbasi, Z., Shekofteh, M., Shahbodaghi, A., & Kazemi, E. (2019). Citation indicators' comparison of LIS open access and subscription publications based on Scopus. *Global Knowledge, Memory and Communication, 68*(4/5), 288–299. https://doi.org/10.1108/GKMC-02-2018-0016

Aguillo, I. F. (2020). Altmetrics of the open access institutional repositories: A webometrics approach. *Scientometrics, 123*(3), 1181–1192. https://doi.org/10.1007/s11192-020-03424-6

Alkhawtani, R. H. M., Kwee, T. C., & Kwee, R. M. (2020). Citation advantage for open access articles in European Radiology. *European Radiology, 30*(1), 482–486. https://doi.org/10.1007/s00330-019-06389-0

Antelman, K. (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

Arendt, J., Peacemaker, B., & Miller, H. (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

Arlitsch, K., & Grant, C. (2018). Why so many repositories? Examining the limitations and possibilities of the institutional repositories landscape. *Journal of Library Administration, 58*(3), 264–281. https://doi.org/10.1080/01930826.2018.1436778

Arlitsch, K., Wheeler, J., Pham, M. T. N., & Parulian, N. N. (2020). An analysis of use and performance data aggregated from 35 institutional repositories. *Online Information Review*. https://doi.org/10.1108/OIR-08-2020-0328

Bautista-Puig, N., Lopez-Illescas, C., de Moya-Anegon, F., Guerrero-Bote, V., & Moed, H. F. (2020). Do journals flipping to gold open access show an OA citation or publication advantage? *Scientometrics, 124*(3), 2551–2575. https://doi.org/10.1007/s11192-020-03546-x

Beatty, J. R. (2019). Revisiting the open access citation advantage for legal scholarship. *Law Library Journal, 111*, 573.

Björk, B.-C., Welling, P., Laakso, M., Majlender, P., Hedlund, T., & Guðnason, G. (2010). Open access to the scientific journal literature: Situation 2009. *PLoS ONE, 5*(6), e11273. https://doi.org/10.1371/journal.pone.0011273

Craig, I., Plume, A., Mcveigh, M., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?A critical review of the literature. *Journal of Informetrics, 1*(3), 239–248. https://doi.org/10.1016/j.joi.2007.04.001

Dorta-González, P., Suárez-Vega, R., & Dorta-González, M. I. (2020). Open access effect on uncitedness: A large-scale study controlling by discipline, source type and visibility. *Scientometrics, 124*(3), 2619–2644. https://doi.org/10.1007/s11192-020-03557-8

Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biology, 4*(5), e157. https://doi.org/10.1371/journal.pbio.0040157

Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies, 1*(2), 618–638. https://doi.org/10.1162/qss_a_00043

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., & Harnad, S. (2010). Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS ONE, 5*(10), e13636. https://doi.org/10.1371/journal.pone.0013636

González-Betancor, S. M., & Dorta-González, P. (2019). Publication modalities ‘article in press’ and ‘open access’ in relation to journal average citation. *Scientometrics, 120*(3), 1209–1223. https://doi.org/10.1007/s11192-019-03156-2

Greene, J. W. (2016). Web robot detection in scholarly Open Access institutional repositories. *Library Hi Tech, 34*(3), 500–520. https://doi.org/10.1108/LHT-04-2016-0048

Greene, J. W. (2017). Developing COUNTER standards to measure the use of Open Access resources. *Qualitative and Quantitative Methods in Libraries, 6*(2), 315–320.

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies, 1*(1), 414–427. https://doi.org/10.1162/qss_a_00022

Lambert, J., & Needham, P. (2019). Institutional repositories and the item and research data metrics landscape. *Insights the UKSG Journal, 32*, 26. https://doi.org/10.1629/uksg.478

Langham-Putrow, A., Bakker, C., & Riegelman, A. (2021). Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PLoS ONE, 16*(6), e0253129. https://doi.org/10.1371/journal.pone.0253129

Lawrence, S. (2001). Free online availability substantially increases a paper’s impact. *Nature, 411*(6837), 521.

McCabe, M. J., & Snyder, C. M. (2014). Identifying the effect of open access on citations using a panel of science journals. *Economic Inquiry, 52*(4), 1284–1300. https://doi.org/10.1111/ecin.12064

Needham, P., & Stone, G. (2012). IRUS-UK: Making scholarly statistics count in UK repositories. *Insights: The UKSG Journal, 25*(3), 262–266. https://doi.org/10.1629/2048-7754.25.3.262

OBrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J., & Borda, S. (2016). Undercounting file downloads from institutional repositories. *Journal of Library Administration, 56*(7), 854–874. https://doi.org/10.1080/01930826.2016.1216224

OBrien, P., Arlitsch, K., Mixter, J., Wheeler, J., & Sterman, L. B. (2017). RAMP – the Repository Analytics and Metrics Portal: A prototype web service that accurately counts item downloads from institutional repositories. *Library Hi Tech, 35*(1), 144–158. https://doi.org/10.1108/LHT-11-2016-0122

Ottaviani, J. (2016). The post-embargo open access citation advantage: It exists (probably), it's modest (usually), and the rich get richer (of course). *PLoS ONE, 11*(8), e0159614. https://doi.org/10.1371/journal.pone.0159614

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ, 6*, e4375. https://doi.org/10.7717/peerj.4375

Razumova, I. K., & Kuznetsov, A. (2019). *Impact of Open Access Models on Citation Metrics*. https://doi.org/10.1633/JISTaP.2019.7.2.2

Xia, J., Lynette Myers, R., & Kay Wilhoite, S. (2011). Multiple open access availability and citation impact. *Journal of Information Science, 37*(1), 19–28. https://doi.org/10.1177/0165551510389358

Young, J. S., & Brandes, P. M. (2020). Green and gold open access citation and interdisciplinary advantage: A bibliometric study of two science journals. *The Journal of Academic Librarianship, 46*(2), 102105. https://doi.org/10.1016/j.acalib.2019.102105

Wheeler, J., & Arlitsch, K. (2021a). *Repository Analytics and Metrics Portal (RAMP) 2017 data* (Version 3, p. 1676180628 bytes). Dryad. Retrieved from https://doi.org/10.5061/DRYAD.R7SQV9SCF

Wheeler, J., & Arlitsch, K. (2021b). *Repository Analytics and Metrics Portal (RAMP) 2018 data* (Version 3, p. 2881804492 bytes). Dryad. Retrieved from https://doi.org/10.5061/DRYAD.FFBG79CVP

Wheeler, J., & Arlitsch, K. (2021c). *Repository Analytics and Metrics Portal (RAMP) 2019 data* (Version 2, p. 2235309805 bytes). Dryad. Retrieved from https://doi.org/10.5061/DRYAD.CRJDFN342

Wheeler, J., & Arlitsch, K. (2021d). *Repository Analytics and Metrics Portal (RAMP) 2020 data* (Version 5, p. 2844233971 bytes). Dryad. Retrieved from https://doi.org/10.5061/DRYAD.DV41NS1Z4

Wheeler, J., & Arlitsch, K. (2021e). *Repository Analytics and Metrics Portal (RAMP) 2021 data* (Version 3, p. 771751265 bytes). Dryad. Retrieved from https://doi.org/10.5061/DRYAD.1RN8PK0TZ

Archambault, É., Amyot, D., Deschamps, P., Nicol, A., Provencher, F., Rebout, L., & Roberge, G. (2014). *Proportion of open access papers published in peer-reviewed journals at the European and world levels—1996–2013*. European Commission. Retrieved from https://science-metrix.com/sites/default/files/science-metrix/publications/d_1.8_sm_ec_dg-rtd_proportion_oa_1996-2013_v11p.pdf

Archambault, É., Cote, G., Struck, B., & Voorons, M. (2016). Research impact of paywalled versus open access papers. *Copyright, Fair Use, Scholarly Communication, Etc*, 6.

Arlitsch, K., & Wheeler, J. (2020). *Repository Analytics and Metrics Portal (RAMP)* [Non-profit]. Retrieved from https://rampanalytics.org

Basson, I. (2019). *An investigation of open access citation advantage through multiple measures and across subject areas for articles published from 2005 to 2014* [Dissertation, Stellenbosch : Stellenbosch University]. Retrieved from http://hdl.handle.net/10019.1/105966

Crossref REST API. (n.d.). Retrieved from https://github.com/CrossRef/rest-api-doc

Gohel, D. (2021). *flextable: Functions for tabular reporting*. Retrieved from https://CRAN.R-project.org/package=flextable

Google. (2020). *Search Console API* (Version 3) [Computer software]. Google. Retrieved from https://developers.google.com/webmaster-tools

Grace-Martin, K. (2012). Can a regression model with a small R-squared be useful. *The Analysis Factor*.

Hajjem, C., Harnad, S., & Gingras, Y. (2006). Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact. *ArXiv Preprint Cs/0606079*.

Hlavac, M. (2018). *stargazer: Well-formatted regression and summary statistics tables* [Manual]. Retrieved from https://CRAN.R-project.org/package=stargazer

Jisc. (nd.). *Sherpa Romeo*. Sherpa Romeo. Retrieved from https://v2.sherpa.ac.uk/romeo/

Jisc. (2022). *OpenDOAR Statistics* [Educational]. OpenDOAR. Retrieved from https://v2.sherpa.ac.uk/view/repository_visualisations/1.html

Macgregor, G. (2019). Improving the discoverability and web impact of open repositories: Techniques and evaluation. *Code4Lib Journal*, *43*. Retrieved from https://journal.code4lib.org/articles/14180

Max Planck Institute. (2003, October 22). *Berlin declaration on open access to knowledge in the sciences and humanities*. Open Access Max Planck Gesellschaft. Retrieved from https://openaccess.mpg.de/Berlin-Declaration

Poynder, R. (2016, September 22). Q&A with CNI's Clifford Lynch: Time to re-think the institutional repository? *Open and Shut?* Retrieved from https://poynder.blogspot.com/2016/09/q-with-cnis-clifford-lynch-time-to-re_22.html

R Core Team. (2020). *R: A language and environment for statistical computing* [Manual]. Retrieved from https://www.R-project.org/

RStudio Team. (2020). *RStudio: Integrated development environment for r* [Manual]. http://www.rstudio.com/

Unpaywall REST API. (2020). Retrieved from https://unpaywall.org/products/api

Wheeler, J., & Arlitsch, K. (2020). *Repository and Analytics Metrics Portal (RAMP) Workflow Documentation and Data Definition.* Retrieved from https://digitalrepository.unm.edu/ulls_fsp/141/

Wheeler, J., Arlitsch, K., Pham, M., & Parulian, N. (2020). *RAMP data subset, January 1 through May 31, 2019* [Data set]. University of New Mexico. https://doi.org/10.5061/dryad.fbg79cnr0