4-2023

Discoverability and Search Engine Visibility of Repository Platforms

Danping Dong
Singapore Management University, dpdong@smu.edu.sg

Aaron Tay
Singapore Management University, aarontay@smu.edu.sg

8. Discoverability and Search Engine Visibility of Repository Platforms

Danping Dong

0000-0002-2229-6709

Chee Hsien Aaron Tay

0000-0003-0159-013X

Abstract

It is of increasing interest and importance to understand the discoverability of content hosted on institutional repositories (IR) and data repositories beyond the library context and on the wider web, as IRs play a significant role in promoting and disseminating the research outputs of an institution. This chapter provides an overview of early studies on IR discoverability in search engines, discusses standardized metrics for repository usage and ways to monitor search engine optimization, and provides practical suggestions to expose IR content to non-Google indexes and aggregators, such as Unpaywall. The second part of this chapter presents a case study and research experiment designed to compare the discoverability and search engine visibility of two hosted repository solutions, Digital Commons and Figshare. No statistically significant differences were found between the two platforms in attracting downloads to deposited academic publications.

Introduction

Scholarly institutional repositories (IR) and data repositories established by higher education institutions or research institutes often have the major goal of managing, disseminating, and providing access to the research outputs produced by their researchers. The institution often highly prioritizes maximizing the visibility and readership of such scholarly content, and search engines are a crucial means to expose repository content to the academic community and a wider audience. This chapter includes a literature review on discoverability of IRs and presents a case study comparing the discoverability and search engine visibility of two hosted repository solutions, Digital Commons and Figshare.

Literature review

For the last decade or two, literature around IRs has focused more on measuring and often bemoaning the lower-than-expected deposit rates at IRs. Until lately, less attention had been given to measuring and improving the discoverability of collected IR content.

Early studies on discoverability of IR content

In the 2010s, studies on discoverability focused on measuring "indexing ratios of IRs in Google Scholar"(Arlistch & OBrien, 2012, 60) or, in other words, the percentage of IR content indexed in Google Scholar. At the time, Arlitsch and OBrien (2012) suggested two methods to determine indexing ratios in Google Scholar. The first method involved searching Google Scholar with the site command of the IR's domain and dividing the number of results by the total number of expected items in the IR that could be found in Google Scholar. This method yielded a shockingly low average of 30 percent. However, Google Scholar does not encourage this method because the site command only shows primary records from an IR. Arlitsch and OBrien's (2012) second method adopted a sampling approach to determine the indexing ratio of IRs by searching the title of each item in Google Scholar. Using this approach, they found three IRs with very high indexing ratios (89–98 percent) while four others had major issues and were indexed at under 50 percent.

Although IR managers might not have expected their items to be highly ranked in the results, being listed in Google Scholar seems to be a fairly low barrier that was still not always met. Many IRs of the time had issues with being properly indexed in Google Scholar. Some of the significant known issues included the following:

- Many IR's support of Dublin Core instead of Google Scholar, which recommended Highwire Press, Eprints, Bepress, and Publishing Requirements for Industry Standard Metadata
- Lack of proper support (Open Archives Initiative Protocol for Metadata Harvesting [OAI-PMH] support was dropped) for standards such as sitemaps to inform Google Scholar of new records
- Poor navigation and a cross-linking structure that deterred Google Scholar bots from indexing the whole repository

However, over the years, members of the Google Scholar team, such as Anurag Acharya, as well as Monica Westin, who was the Google Scholar Partnerships Lead, have been communicating more with IR developers and managers, and are providing guidance on best practices, common pitfalls, and fixes for common indexing issues (Acharya, 2015; Westin, 2019; Westin, 2021). These efforts have helped improve discoverability of IRs as measured by their indexing ratios. This is particularly true for popular, well-established IR systems like DSpace, Eprints, and Digital Commons which have worked with Google Scholar over the years to ensure reasonable out-of-the-box settings that have resolved many problems that would have previously resulted in low indexing ratios (COAR, 2018). Still, it is highly recommended to check for best practices with peers or the platform vendor when setting up an IR. It is also equally important to conduct periodic checks. Westin's (2021) recent talk could serve as a guide for beginners. Furthermore, the checklists for DSpace repositories (EIFL, 2021) and COAR Repository Toolkit (COAR, 2018) will also prove helpful.

Metrics for measuring discoverability of IRs

To begin measuring discoverability of IRs, reliable metrics are needed for comparison. Arlitsch & Obrien's 2015 paper and toolkit is important in the area of measuring discoverability in IRs. It provides a step-by-step guide on how libraries can measure and monitor the search engine optimization (SEO) of IRs. The guide advocated monitoring not just Google Analytics but also Google Webmaster Tools, now known as Google Search Console.

Obrien et al. (2016) note that there is a need to combine web analytics metrics from both page tagging analytics (e.g., Google Analytics) and log file analytics (e.g., build-in packages built into DSpace and Eprints, etc.). However, both methods have risks in terms of over- or under-counting metrics such as visits, downloads, and page views. While web analytics from page tagging are popular and typically easy to analyze, they can only capture HTML views or page loads. This is problematic because the major source of traffic to IR content is visits to the PDFs of resources directly via Google Scholar. In particular, they showed that relying solely on Google Analytics seriously under-counts file downloads when using page views as a proxy.

On the other hand, while log analysis can track all the content that users have interacted with, rather than only HTML, it has traditionally been more difficult to analyze (though improvements in technology have reduced this issue). More seriously, log analysis metrics are prone to over-counting because it is difficult to distinguish bots and web scrapers that hit the servers. Some estimate as much as 30–50 percent of log traffic comes from bots.

Therefore, it is clear that to get a complete picture of the analytics of IRs, we will need to use a hybrid system. One common and highly popular solution is leveraging the use of the free Google Analytics and Google Search console tools together with an IR's logging features, though the use of Google products comes with privacy concerns.

In particular, Google Search Console is powerful as it allows the tracking of every page-click to the handle (aka URL) of every repository item that appears in the Google Search Engine Result Page. In other words, it "provides accurate non-HTML download counts executed directly from all Google search engine results pages (SERP)" (Arlitsch et al., 2020, p. 317), which avoids the issue of Google Analytics not capturing such downloads.

Standardizing metrics across institutions – RAMP, IRUS-UK – and the Making Data Count project

As IRs started to band together to compare notes, it was natural to consider whether the metrics collected across different repositories could be standardized for benchmarking purposes. For example, when comparing log analysis results, it would be beneficial to have the same procedure and lists of bots to be filtered to allow metrics collected around different repositories to be comparable. If repositories were using Google Analytics and/or Google Search Console, using the same parameters for comparison was beneficial.

There are two major approaches to benchmarking repository metrics. First, in 2017, the Repositories Analytics & Metrics Portal (RAMP) was launched by Michigan State University

Library and partners such as OCLC Research and Association of Research Libraries, with Institute of Museum and Library Services research funding (Arlitsch et al., 2018; OBrien et al., 2017). RAMP (https://rampanalytics.org/) is a web service to collect, standardize, and analyze Google Search Console data from participating repositories with data aggregated in an Elasticsearch index. Each registered institution is given access to a set of dashboards on Kibana. At present, over 60 IRs from multiple countries are included.

Second, we have the Institutional Repository Usage Statistics UK (IRUS-UK) group run by JISC in the UK. While RAMP focuses on aggregating and standardizing Google Search Console data, IRUS-UK focuses on handling log data files and processing them into a standardized Counting Online Usage of Networked Electronic Resources (COUNTER) statistics (MacIntyre & Jones, 2016). Similar to RAMP, once registration into IRUS-UK is done and data is being sent, usage can be checked via the web portal, which provides various reports and visualizations at institutional or item level. IRs like Eprints, Figshare, Pure, and DSpace have tracker code plugins and patches that can help submit data to IRUS-UK.

While standardized COUNTER statistics techniques can be easily applied to articles, IRs today also collect research data. Is there a COUNTER standard for research datasets?

This is where the Making Data Count project comes in, which introduced the COUNTER Code of Practice for Research Data in 2018. While research dataset usage can be defined easily as downloads or views (at the file or dataset level), the lack of standardized definitions means that the tracking and display of such statistics by repositories and stakeholders is, to some extent, arbitrary. As Lowenberg et al. (2019) put it, "Currently, to compare the downloads across datasets within a repository, or across repositories, would be comparing apples to oranges, as we do not know where these numbers are derived from, nor exactly what they apply to" (p. 30). A standard that is not adopted is not helpful. At the time of writing, uptake of the COUNTER Code of Practice for Research Data is promising. Data repositories, repository systems, and aggregator organizations who have adopted this include Figshare, Dryad, Zenodo, Dataverse, DataONE, and Caltech.

Google Dataset Search – the new kid on the block

While most repository managers are familiar with Google Scholar and Google, in 2018 Google added a new search engine exclusively for datasets, The Google Dataset Search, which quickly came out of beta in January 2020, and which has since attracted considerable interest.

Without going into specific details, the Google Dataset Search bot crawls webpages looking for Schema.org markup to index. This is also stated as a requirement for repositories to be listed on it. Similar to Google Scholar, a sitemap is not strictly required but highly recommended.

Do most data repository systems support this? " A data citation roadmap for scholarly data repositories" report listed the following requirement as *recommended* : "The machine-readable metadata should use schema.org markup in JSON-LD format" (Fenner et al., 2019, Table 1 Guidelines for Repositories.) However, with the launch of Google Dataset Search, many data

repository vendors such as Figshare, Mendeley Data, Zenodo, and Dryad support this so as to be listed in Google Dataset Search.

Because this development was relatively recent at the time of writing, there are limited guidelines (https://datasetsearch.research.google.com/help) and research on the discoverability of repository items in Google Dataset Search, although it is likely that many of the usual Google Scholar techniques work. Sampling datasets in Google Dataset Search may be a worthwhile exercise to ensure that most datasets are properly indexed, and to use markup tools to verify the quality of metadata in the markup. The FAQ for Google Dataset Search offers more details by discussing the use of Markup Helper to create metadata and the Structured Data Testing Tool to verify if the metadata is correct. Unfortunately, while many data repositories, such as Figshare, Dryad, Dataverse, and Mendeley Data, support Schema.org (see list here), many others may not. This is bad, particularly as there is increasing evidence that supporting Schema.org might increase discoverability.

What if you have a data repository but do not have the capability to add Schema.org to the landing pages? Are you completely invisible? Not quite. Chances are you registered your dataset with a DataCite DOI, and DataCite has done some work to ensure their entry is indexed in Google. As datacite.org states, "If a data repository doesn't provide schema.org metadata via the dataset landing page, the next best option is the indexers that store metadata about the dataset. DataCite Search is such a place, and in early 2017 we started to embed schema.org metadata in DataCite Search pages for individual DOIs, and we generated a sitemaps file (or rather files) for the over 10 million DOIs we have" (Cousijn et al., 2018).

Unpaywall and other non-Google channels

While Google, Google Scholar, and Google Dataset Search are the major sources of visitors, there might be a need to ensure that discoverability in other aggregators, indexes, and search engines is not neglected. Before doing so, the low hanging fruit is to ensure that the repository is listed in registries of repositories whenever possible, as this is where aggregators generally start. At the time of writing, three major registries exist:

- OpenDOAR (https://v2.sherpa.ac.uk/opendoar/)
- Registry of Open Access Repositories (http://roar.eprints.org/)
- Registry of Research Data Repositories (https://www.re3data.org/)

Some sources worth checking for discoverability of content are as follows:

- CORE (https://core.ac.uk/)
- BASE (https://www.base-search.net/)
- OpenAIRE (https://www.openaire.eu/)
- Unpaywall (https://unpaywall.org/)
- OpenAlex (https://openalex.org/)

Other sources that the University of Liège Library used to optimize their repository to make it more discoverable (Bastin & Renaville, 2018) include the following:

- PubMed LinkOut
- Primo Central Index
- Summon
- EBSCO Discovery Service

Of the sources above, Unpaywall may be one of the most important non-Google sources to ensure the discoverability of repository content, as it has become the de facto source that most abstracting and indexing databases, search engines, and link resolvers use to link to open access (OA) journal articles. Some of the important consumers of Unpaywall data include the following:

- Web of Science
- Scopus
- Dimensions
- SFX
- 360Link
- Primo link resolver

Besides driving repository usage, Unpaywall is also an important source of OA data for research studies and university rankings, such as the Centrum voor Wetenschap en Technologische Studies Leiden Ranking. Given the ease of querying Unpaywall to check OA status, most research studies use Unpaywall data to determine the percentage of university output that is made OA. If such results are important, ensuring high discoverability of repository content with quality metadata in Unpaywall is also important.

If an IR is not yet indexed in Unpaywall, a request to be indexed (https://unpaywall.org/sources) can be submitted. Like Google Scholar, the Unpaywall bot may sometimes have issues while fully indexing repository content. According to the support page (https://support.unpaywall.org/support/solutions/articles/44001937113-how-are-documents-located-from-repository-records-), Unpaywall uses OAI-PMH to identify records and attempts to look for a URL leading to a record. Depending on the repository setup, the URL is often not the PDF but a landing html page, and from there the Unpaywall bot will try to identify the direct link to the full text, typically a PDF or an html page. Unpaywall is often *smart* enough to identify the full-text link, although it is not 100 percent reliable.

Unpaywall also needs two pieces of metadata information that is sometimes hard to come by in typical repository setups:

- Version of full text
- Usage license

When testing your repository for discoverability in Unpaywall, consider doing the following:

1. Do an internal calculation on the percentage of repository records with full text
2. Run the Unpaywall application programming interface over your repository records (using DOIs as a priority followed by title) to calculate the same metric based on Unpaywall data

3. Compare (1) and (2); if they are significantly different, your repository may have a problem with indexing by Unpaywall

In particular, the percentage of (2) can be lower than (1) because it indicates that either the record itself is not indexed by Unpaywall, or the metadata record is indexed but Unpaywall is unable to locate the full text. For better accuracy of OA status on Unpaywall, it is also important to check the field reflecting the OA status (some IRs use the license field) and to ensure that the field is properly captured by Unpaywall. Working with Unpaywall support to resolve any existing issues may be required.

Repository managers should also be familiar with the following Unpaywall support pages (https://support.unpaywall.org/support/solutions/folders/44000583618); a few important ones include the following:

- Recommendation for IRs: Version reporting https://support.unpaywall.org/support/solutions/articles/44000826872-recommendation-for-irs-version-reporting
- How are repository records matched to published articles? https://support.unpaywall.org/support/solutions/articles/44001937102-how-are-repository-records-matched-to-published-articles-
- How are documents located from repository records? https://support.unpaywall.org/support/solutions/articles/44001937113-how-are-documents-located-from-repository-records-
- Recommendation for IRs: License reporting https://support.unpaywall.org/support/solutions/articles/44002198169-recommendation-for-irs-license-reporting

Discoverability and SEO of hosted solutions

Chapter 2 discussed several features and functionalities affecting discoverability of content from outside of the repository platform, including metadata management, support of harvesting standards and protocols, as well as SEO. Certain adjustments and improvements can be made to a repository to improve its discoverability (Macgregor, 2019). For tweaking and adjusting repository settings to improve discoverability, open source solutions often offer more flexibility as changes can be made when needed. In comparison, it might be less straightforward, or sometimes not possible, to customize certain technical aspects of a proprietary hosted solution, hence the need to thoroughly investigate its discoverability and SEO before adoption. Whether to run an open source or proprietary repository solution depends on the institutional context. Institutions that adopt an open source solution often need to hire staff with the relevant technical expertise to implement and maintain the repository software, whereas institutions with proprietary solutions may often place more emphasis on day-to-day operations, growing content, and supporting faculty services, leaving the technical aspects to the vendor. For adopters of hosted and proprietary solutions, it is important to make sure that the right choice is made at the beginning.

There has been little prior research attempting to answer the question of whether repository platforms inherently differ in terms of their discoverability and SEO. Comparing existing repositories hosted on different platforms is not helpful; it will neither be a fair comparison nor statistically convincing as a post-hoc analysis that does not control for or isolate other factors that impact usage and discoverability, such as quality and age of the papers and subject disciplines.

The case study in the next section of this chapter outlines a method that uses a randomized controlled trial (RCT) to compare two hosted solutions, Digital Commons and Figshare. It attempts to examine whether these two platforms differ in their ability to attract downloads to hosted papers. It also explores patterns of paper indexing and visibility in Google Scholar. The method as well as the results may provide some useful information for institutions that are planning to adopt a hosted IR or are considering switching to another platform.

Case study: Comparing discoverability of Digital Commons and Figshare

We conducted an RCT to compare the downloads of two hosted IR solutions, Digital Commons and Figshare, and explored the patterns used for indexing their records in Google Scholar.

This case study is conducted using the IR and data repository of Singapore Management University (SMU), hosted on Digital Commons and Figshare, respectively. SMU has been running its IR, InK (https://ink.library.smu.edu.sg/), on Digital Commons since 2011. In April 2020, SMU started its Research Data Repository (RDR) (https://researchdata.smu.edu.sg/) with the hosted solution Figshare. While Figshare is more well-known as a research data repository, some institutions also use Figshare as an all-in-one repository for mixed types of research outputs, including publications, for example, Carnegie Mellon University's KiltHub Repository (https://kilthub.cmu.edu/). While this is a possible direction for SMU to move toward in the future, we do not plan to rush migrating to a new repository and risk losing the existing advantages of Digital Commons, such as good discoverability, comprehensive statistics, and institutional users' familiarity and acceptance.

In 2021, we conducted an exploratory project to study the feasibility of using Figshare as an IR and to compare it with Digital Commons. One major objective of this project was to understand the discoverability and search engine visibility of the repository. Therefore, we decided to conduct an experiment to study whether the downloads of deposited papers will be impacted by the IR platform used.

Methodology

Hypothesis development

The main purpose of this study was to explore whether any platform difference exists between Digital Commons and Figshare in attracting downloads to deposited academic publications. We planned to experiment with two random groups of full-text journal articles uploaded to both platforms around the same time. The usage and download statistics of both groups were tracked and monitored over seven months. We worked with the assumption that other factors affecting

downloads, such as quality of the article and popularity of research topics, will be randomized among the two groups. Therefore, the difference in download counts can serve as a reasonable approximation of the platform discoverability difference between Digital Commons and Figshare.

We established the following hypothesis:

> **H0: There is no difference in average paper downloads between Figshare and Digital Commons.**
>
> **H1: Average paper downloads differ between Figshare and Digital Commons.**

The significance level was set to be 0.05.

Setting up test collections of randomly assigned full-text records

To compare the platform discoverability of Digital Commons and Figshare, we decided to use download counts, which is a reasonable measure of usage resulting from web traffic driven to the repository. We set up two test collections with randomly selected journal articles on both platforms. The source of the articles was SMU's Current Research Information System (CRIS), which is used by our faculty to update their publications for reporting purposes. The disciplinary coverage of the articles thus included social sciences, business, computing, law, economy, and accountancy, contributed by faculty from the six schools of SMU.

A total of 96 journal article records with full-text PDFs were exported from CRIS. Half the records (48) were uploaded to InK, SMU's IR hosted on Digital Commons, and the other half to SMU RDR on Figshare. As seen from Figure 8.1, 79.3 percent of the articles were published recently in 2020 and 2021, which is expected as most faculty use CRIS to report on their latest academic publications.

For the rest of this article, Digital Commons will be used to refer to InK records, and Figshare will be used for records uploaded to RDR.

<Fig. 8.1 here>

Figure 8.1 Distribution of article publication year in the sample (by count).

We also attempted to control for the assignment of articles by discipline, a factor known to affect usage metrics such as citations (Harzing, 2016; Marx & Bornmann, 2015). As mentioned earlier, the SMU schools associated with the records can serve as an approximation of the research areas of the articles, that is, social sciences, business, computing and information systems, law, economy, and accountancy. Within each research area, records were randomly divided by half and assigned to Digital Commons and Figshare.

Data collection

The journal article metadata and full-text PDF were uploaded to both platforms toward the end of March 2021. Monthly download count statistics of each article from both platforms were collected from April to October 2021. The default download counts in Figshare include bots downloads, which are excluded in the statistics from Digital Commons. We therefore requested download counts excluding bots downloads from Figshare, who readily provided this information.

During the course of the study, a few records were removed because of issues such as duplication or author request. The final dataset as of December 2021 contained 45 valid records from Digital Commons and 47 from Figshare.

In June 2021, we discovered that records from Digital Commons were not properly indexed by Google Scholar, which might affect the discoverability of Figshare and our ability to perform a fair comparison between the two platforms. We approached Figshare support and were advised to fix issues with the field used for publication dates, which may have caused Google Scholar to ignore the Figshare records. After fixing the issue, we noted that the records appeared in Google Scholar around the end of July. To minimize the impact of this incident, we will analyze both the full dataset and a subset of the collected data from August 2021 onward.

In addition to testing the proposed hypothesis, we were also interested in exploring how records from our repositories were indexed in Google Scholar. In late September, we did a round of checking and data collection about Google Scholar indexing and added three additional fields to the dataset. We searched for each article by title and checked the following:

1) Whether the record is indexed in Google Scholar at all
2) If the record is in Google Scholar, whether our record is the only copy providing a unique PDF among the different versions of the same article after clicking “All n versions”
3) Whether our record is shown as the primary record in Google Scholar (see Figure 8.2)

<Fig. 8.2 here>

Figure 8.2. Example of a primary record from Figshare on Google Scholar.

Results

For this research, we used a two-sample t-test and a few other statistical tests, such as Levene’s test, to assess our hypothesis and analyze the collected data. The analysis was conducted using Python in Jupyter Notebook, and the analysis can be reproduced with the raw data and code publicly shared on SMU RDR (Dong & Tay, 2022).

Hypothesis testing

Some descriptive statistics are included for the 92 records in Table 1 below.

Name of IR	count	mean	std	min	25 %	50 %	75 %	max
InK (Digital Commons)	45.0	51.311111	78.227472	1.0	8.0	23.0	47.0	354.0
RDR (Figshare)	47.0	64.021277	142.855885	0.0	6.0	26.0	59.0	934.0

Table 8.1 Descriptive statistics of Digital Commons versus Figshare downloads

According to Table 8.1, the mean download count of Digital Commons is 51 and that of Figshare is 64. We used Levene's test to assess the equality of variances.

LeveneResult(statistic=0.3513837444070336, pvalue=0.5548174249693081)

The resulting p-value = 0.55 implies that the variances between the two groups are not significant. We then used a two-sided standard independent two-sample t-test that assumes equal variance to test our null hypothesis that there is no difference in average paper downloads between Figshare and Digital Commons. The result of the t-test is included below.

Ttest_indResult(statistic=-0.5260135324877442, pvalue=0.600172599662083)

The p-value is 0.596 and greater than the specified significance level of 0.05. Therefore, we failed to reject H1, which leads to the conclusion that there is no statistically significant difference in the download counts between the two platforms.

Additional analysis related to Google Scholar

We conducted further analysis on the data collected about Google Scholar to explore and gain insights on the discoverability of records in search engines, and to observe if there are any patterns and relationships between how records are indexed and the associated downloads. The analysis and results are presented below, providing some answers and insights to the three questions raised earlier.

1) Comparing download counts for records indexed versus not indexed by Google Scholar
   All 45 Digital Commons records are indexed by Google Scholar, reaching a 100 percent indexing rate. However, we noticed that 20 percent (n = 9) of the Digital Commons records do not appear with a [PDF] icon next to the record, even though full-text PDFs are available in the repository. These nine records neither provide a unique PDF nor are they the primary

records. There are alternate OA copies on Google Scholar; one has to click the "All n versions" button, and the Digital Commons record is usually not prominent on that page.

On the other hand, 68 percent of the 47 Figshare records are indexed by Google Scholar as of September 2021. The records that do not show up on Google Scholar can still be found in Google, which means that the download traffic observed for these articles might be mainly from Google. There are no obvious patterns observed in these non-indexed articles. Almost all of them can be found in Google Scholar with alternative sources of metadata or full text. There is no clear reason why our Figshare full-text records are picked up by Google but not by Google Scholar.

As there is no statistically significant difference between download counts of Digital Commons and Figshare, we do not differentiate by platform when comparing downloads between indexed and non-indexed records. Due to the Google Scholar indexing issue for Figshare records described earlier in this chapter, we decided to only include data from August to October 2021 for a fair comparison. Similar to the earlier analysis, we used Levene's test for equality of variance and t-test for comparing downloads. We also performed the Mann–Whitney U test as the sample size of non-indexed records was relatively small (n = 16). The results are presented below:

LeveneResult(statistic=0.4149103399346608, pvalue=0.521125917434428)
Ttest_indResult(statistic=0.5681030671982966, pvalue=0.5713796493618881)
MannwhitneyuResult(statistic=581.5, pvalue=0.9704388586933024)

Based on the results, there is no observed difference in the variance of indexed v non-indexed records. Both the t-test and Mann–Whitney U test show that there is no statistically significant difference for the download counts of indexed versus non-indexed records in Google Scholar. This is a surprising result that we will discuss later in this chapter.

2) Compare records that provide a unique PDF in Google Scholar versus those that do not
Another question explored was whether a record that provides the only full-text PDF in Google Scholar is correlated with higher downloads. As seen in Table 8.2, there are a total of 77 records indexed by Google Scholar, and 44.2 percent (n = 34) of the records provide the only full-text PDF in Google Scholar. In addition, 88.2 percent (n = 30) of them did appear as primary records, which is somewhat expected because Google Scholar is likely to prioritize the ranking of such records in the algorithm.

	Count	mean	Std	min	25%	50%	75%	max
uniq_PDF								
False	43.0	31.4	47.8	0.0	6.00	16.0	35.50	257.0

	Count	mean	Std	min	25%	50%	75%	max
uniq_PDF								
True	34.0	98.8	173.5	4.0	16.75	40.0	90.75	934.0

Table 8.2. Descriptive statistics of download counts for records with and without a unique PDF

We conducted Levene's test and found that the variances between the two groups are not equal. Subsequently, we used a one-sided t-test to test the hypothesis that records with unique PDFs are correlated with higher downloads. The results confirmed our hypothesis with a p-value less than 0.05. The mean download count of records with unique PDFs (98.8/article) is 2.15 times higher compared to records with non-unique PDFs (31.4/article).

LeveneResult(statistic=4.595255876125392, pvalue=0.03530232522808095)
Ttest_indResult(statistic=2.2001087678993465, pvalue=0.017063247719817036)

From the results, it can be inferred that IR records that provide a unique PDF in Google Scholar are positively correlated with a higher number of downloads compared to records with full-text PDFs that were already available elsewhere.

3) Compare primary versus non-primary records in Google Scholar

We also speculated that when we have records that show up as primary records in Google Scholar, these records should be able to get more downloads compared to non-primary records, since the primary records get more exposure and visibility when searching. Therefore, we compared the download counts of these two groups, and the descriptive statistics are listed in Table 8.3. Three records were excluded from the analysis due to inaccurate primary links, owing to a technical issue from Figshare during the course of the study.

	count	mean	std	min	25%	50%	75%	Max
primary								
FALSE	22.0	12.1	11.4	0.0	4.00	7.5	21.25	42.0
TRUE	52.0	82.5	146.6	1.0	13.75	37.5	79.50	934.0
not sure	3.0	51.7	15.0	36.0	44.50	53.0	59.50	66.0

Table 8.3. Descriptive statistics for download counts of primary versus non-primary records

As seen from Table 8.3, the mean download count of primary records (n = 52) was 82.5 per article, which is 5.8 times higher compared to non-primary records (n = 22). We used Levene's test and a one-sided t-test to show that the two groups do not have equal variance, and that the download count of primary records is higher than that of non-primary records. Test results are indicated below.

LeveneResult(statistic=4.083484886906183, pvalue=0.04702404022812365)
Ttest_indResult(statistic=3.437979553995462, pvalue=0.0005779416388394055)

Discussion

1. The two platforms do not differ significantly in terms of attracting paper downloads

The main research question of this study is whether there is any platform-level difference in terms of exposing OA content and attracting downloads between Digital Commons and Figshare. Based on the results of the RCT study, the two platforms do not differ significantly in their ability to expose IR content and attract paper downloads.

As repository managers, it is important to assess platform discoverability when selecting a hosted IR solution. One of the major goals of IRs is often to increase the visibility and readership of an institution's research outputs to reach a wider audience; the number of paper downloads is an important indicator and metric to measure how successful an IR is in achieving this goal. Institutions that would like to start an IR or would like to move to another platform should thoroughly investigate platform discoverability and SEO by performing literature searches, investigating existing repositories, or if circumstances allow, conducting a pilot test or comparison study such as the one outlined in this paper.

Both platforms are satisfactory in terms of their discoverability and content usage. We have been using the Digital Commons platform as our IR for more than 10 years and are happy with the usage reaching a total of 5.9 million downloads (as of Nov 2021). Digital Commons is considered to be a mature hosted IR solution, having existed for 17 years, and is claimed to be optimized for indexing by Google, Google Scholar, and other major search engines. Figshare was launched in 2011 and is more well-known as a research data repository focusing on non-publication outputs. From 2017, Figshare gradually added more IR functionality and started describing itself as an "all in one repository" (Hyndman, 2018). Interestingly, the downloads of journal article content type on Figshare are at a similar level with a more established platform such as Digital Commons, confirming Figshare's potential to serve as a decent publications repository in addition to being a data repository.

Nevertheless, some differences exist between the platforms pertaining to Google Scholar indexing. Digital Commons achieved a 100 percent indexing rate in our study, although some of the full-text records are not correctly indicated as such on the platform, and direct links to the PDFs are not shown to Google Scholar users. In comparison, the indexing ratio of Figshare is only 68 percent and is therefore lower than expected. Although this did not result in a significant difference in downloads in our experiment, there might be other downstream implications. It is reasonable to

expect that researchers might be concerned if their papers do not appear on Google Scholar, which likely caters to a more academic audience who is more likely to cite these papers.

We speculate that the indexing ratio for Figshare records is lower because after the initial publication date issue with Google Scholar was fixed, the re-indexing of our Figshare repository was partial and incomplete. Given more time, some of the records might eventually appear on Google Scholar. If this is the case, our poor indexing ratio should be regarded with a pinch of salt. Further investigation, such as repeating the experiment with another batch upload of test records on Figshare, could be done to measure the indexing ratio again to rule out this possibility. It is also possible that some other reasons innate to the platform might have caused the relatively lower Google Scholar indexing ratio. Therefore, based on the results from this experiment so far, Digital Commons seems to be more consistent and reliable in terms of exposing its content to Google Scholar.

2. Google versus Google Scholar referrals

Based on our earlier analysis, Google Scholar indexed versus non-indexed records do not differ significantly in terms of the number of downloads. The result is also in agreement with our observation during the experiment itself. However, from April to July, there was a problem with Google Scholar indexing for Figshare records. Even though we excluded data from this period from the analysis, we observed that the downloads during that period do not seem to be affected much, if at all.

We also obtained the referral statistics provided by Digital Commons for all time and calculated the percentage of downloads contributed by Google.com and Google Scholar (see Table 8.4).

	No. of referred downloads	**% of contributed downloads**
Google only	2845405	61%
Google Scholar only	1210470	26%
Google & Google Scholar	4055875	88%
InK	336782	7%
Total	4631336	

Table 8.4 Referral analysis of InK downloads since launch

Table 8.4 shows that Google.com alone contributed to 61 percent of total download traffic for our InK repository, considerably higher than the download referrals from Google Scholar (26 percent). Both search engines under Google contributed to 88 percent of our entire site's traffic, clearly demonstrating the crucial impact of search engine traffic on the usage of IRs. In comparison, only 7 percent of the downloads occur on the IR site itself. This aligns with results from another study on Strathprints, the University of Strathclyde repository, which found that 56 percent of all referrals came from Google, followed by 26 percent from Google Scholar (Macgregor, 2020).

Although the results from our experiment seem to suggest that being indexed in Google Scholar is not a significant factor correlated with higher download counts, we should not underestimate the importance of Google Scholar for IRs. First, it could be argued that traffic that occurred in Google Scholar may have a higher chance of leading to a future citation, as more academic research-related searches are likely to happen in Google Scholar than in Google. From the perspective of an IR manager, we should emphasize exposing our institution's content to the research community, and the fact that Google Scholar is a vital channel to achieve this goal. In addition, Google Scholar itself also contributed significantly to download traffic, as seen from InK's historical data for all our repository records.

Furthermore, repository managers should not assume proper indexing of all repository records by Google Scholar, especially when starting a new repository. Configuration of publication dates on our Figshare repository was an unexpected issue that resulted in a Google Scholar indexing problem. Repository managers should therefore take special effort to check and track the indexing of records in Google Scholar.

3. Further analysis of unique and primary records

Our exploratory analysis earlier suggested that records providing a unique PDF, or records that display as the primary records, are both likely to receive higher downloads. It is reasonable that primary records are likely to get higher downloads since they are more prominent when people search in Google Scholar. It is also plausible that being the unique OA source for a paper increases the likelihood for a user to eventually download the article from the repository, compared to when the IR needs to compete with other OA services and repositories such as ResearchGate, SSRN, publishers, and Arxiv.org. Moreover, the fact that a record is the unique source of an OA full text in Google Scholar also means that it is highly possible that it is the only OA copy on Google as well. All these factors might have contributed to the higher downloads for unique or primary records.

We also speculate that being a unique OA source is linked to a higher likelihood of being indexed by Google Scholar as a primary record. Based on Table 8.5, 90.6 percent of the records containing a unique full text in Google Scholar are listed as the primary record, compared to 54.8 percent for non-unique OA copies. Among the three copies that were unique but non-primary, one was not a primary record at the time of checking (September 2021) but has become a primary record since November 2021. The primary record of the other two articles both point to the publisher's official landing page with the "[HTML]" icon and, in both cases, lead to the publisher's OA version of that article. One of the articles was found to be published in a fully OA journal, while the other was a gold OA article published in a hybrid journal. It is possible that Google Scholar was able to identify these two articles as publisher OA content, thus prioritizing them as primary records. That also means that our IR record in this case does not provide a unique OA copy even though it is the only record with a "[PDF]" button in Google Scholar.

	Unique OA copy	

Primary record	FALSE	TRUE	Total
FALSE	19	3	22
TRUE	23	29	52
Total	**42**	**32**	**74**

Table 8.5 Matrix of Google Scholar unique OA copy against primary record

The exact requirements and mechanism of becoming a primary record are unclear since Google does not make its algorithms transparent. However, providing a unique OA full text is more in the control of a repository. Theoretically, if an IR undertakes greater conscious effort in collecting unique OA full text that is not available elsewhere, it might help to attract more traffic to the repository. Realistically, most IRs will likely collect full text – whether or not there are already OA copies elsewhere – but it is still worth noting when planning and strategizing the development and growth of an IR.

We also closely analyzed records that do not provide a unique PDF in Google Scholar from Digital Commons (n = 29) and Figshare (n = 14). Among the non-unique records in Digital Commons, 65.5 percent (n = 19) appear as the primary record in Google Scholar, while 28.6 percent (n = 4) of the non-unique records in Figshare became the primary record. While this observed pattern could be purely incidental and possibly unrelated to the platform's intrinsic characteristics, it is interesting to explore whether platforms may be optimized so that there is a higher chance of their records being indexed by Google Scholar as primary records.

Limitations of this study

One limitation of this study is the Figshare indexing issue by Google Scholar that occurred from April to June 2021. This was an unexpected episode that may or may not have impacted the results of this study. It should not have much impact on the main research question studying discoverability of these two platforms, as our focus was on overall downloads to deposited papers and not specifically those from Google Scholar. Download patterns do not seem to differ before and after fixing the issue, and Google Scholar indexing was not correlated with higher downloads based on our results. Nevertheless, we are not sure if this issue had any impact on the lower-than-expected indexing rate of Google Scholar in our exploratory analysis.

Conclusion

By conducting an RCT of papers deposited in Digital Commons and Figshare, we were able to perform a fair comparison of their relative discoverability in terms of downloads to deposited papers. We concluded that there is no evidence supporting a platform-level difference. However, further exploratory analysis revealed a notable difference in their Google Scholar indexing ratio, yet this did not seem to affect the overall download counts. Our results, which are supported by prior literature, suggest that overall downloads might be heavily influenced by Google rather than Google Scholar, though further study is needed for confirmation. Our exploratory findings also

suggest that records that are unique in Google Scholar or those that are listed as primary records tend to be associated with more downloads on average. However, due to a lack of controls, these results are preliminary. Future research can focus on the following:

1) Why are certain items deposited in Figshare not indexed by Google Scholar?

2) Attempt to unravel the relative importance of downloads from Google versus Google Scholar through the use of Google Search Console, Google Analytics, and log analysis (Obrien et al., 2016), as well as the possible long-term impact on citations for items not indexed in Google Scholar.

References

Acharya, A. (2015, June 10). *Indexing repositories: Pitfalls and best practices—Media collections online*. https://media.dlib.indiana.edu/media_objects/9z903008w

Arlitsch, K., Kahanda, I., OBrien, P., Shanks, J. D., & Wheeler, J. (2018). *Data-driven improvement to institutional repository discoverability and use*. https://scholarworks.montana.edu/xmlui/handle/1/15631

Arlitsch, K., & OBrien, P. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, 30(1), 60–81.

Arlitsch, K., & OBrien, P. (2015). Introducing the "getting found" web analytics cookbook for monitoring search engine optimization of digital repositories. *Qualitative and Quantitative Methods in Libraries (QQML)*, *4*, 947–953.

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*, *45*(2), 316–335. https://doi.org/10.1108/OIR-08-2020-0328

Bastin, M., & Renaville, F. (2018). *Open access discovery: ULiège experience with aggregators and discovery tools providers. Be proactive and apply best practices (if you can...)*. https://orbi.uliege.be/handle/2268/221340

COAR. (2018). *COAR repository toolkit*. https://coartraining.gitbook.io/coar-repository-toolkit/

Cousijn, H., Cruse, T., & Lammey, R. (2018, September 5). Taking discoverability to the next level: Datasets with DataCite DOIs can now be found through Google Dataset Search. *DataCite Blog*. https://blog.datacite.org/taking-discoverability-to-the-next-level/

Dong, D., & Tay Chee Hsien, A. (2022). *Data and code for the case study comparing discoverability of Digital Commons and Figshare*. https://doi.org/10.25440/smu.19121768

EIFL. (2021). *EIFL Checklist: How to make your OA repository work really well (Version 5) | EIFL*. https://www.eifl.net/resources/eifl-checklist-how-make-your-oa-repository-work-really-well-version-5

Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., & Clark, T. (2019). A data citation roadmap for scholarly data repositories. Scientific Data, 6(1), 28. https://doi.org/10.1038/s41597-019-0031-8

Harzing, A.-W. (2016). *Citation analysis across disciplines: The impact of different data sources and citation metrics*. https://harzing.com/publications/white-papers/citation-analysis-across-disciplines

Hyndman, A. (2018, December 11). *New funding information on Figshare items*. https://figshare.com/blog/New_funding_information_on_Figshare_items/446

Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., & Jones, M. B. (2019). *Open data metrics: Lighting the fire (Version 1)* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3525349

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

Macgregor, G. (2020). Enhancing content discovery of open repositories: An analytics-based evaluation of repository optimizations. *Publications*, *8*(1), 8. https://doi.org/10.3390/publications8010008

MacIntyre, R., & Jones, H. (2016). IRUS-UK: Improving understanding of the value and impact of institutional repositories. *The Serials Librarian*, *70*(1–4), 100–105. https://doi.org/10.1080/0361526X.2016.1148423

Marx, W., & Bornmann, L. (2015). On the causes of subject-specific citation rates in Web of Science. *Scientometrics*, *102*(2), 1823–1827. https://doi.org/10.1007/s11192-014-1499-9

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

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

Westin, M. (2019, December). *DSpace and Google Scholar Webinar for Ghana*.

Westin, M. (2021, January 27). *Google Scholar Indexing for Repositories: Best Practices and Fixes for Common Indexing Problems*. https://www.youtube.com/watch?v=C-miRaROsaE