Promoting content discovery of open repositories : reviewing the impact of optimization techniques (2016-2019)

George Macgregor
https://purl.org/g3om4c [1]

[1]Scholarly Publications & Research Data Team, IS Information Management, University of Strathclyde, Glasgow, Scotland UK

2019-05-20

Abstract

Ensuring open repositories fulfil the discovery needs of both human and machine users is of growing importance and essential to validate the continued relevance of open repositories to users, and as nodes within open scholarly communication infrastructure. Following positive preliminary results reported elsewhere, this submission reviews the longer term impact of a series of discovery optimization approaches deployed on an open institutional repository. These approaches were designed to support improved content discovery and user engagement, thereby improving content usage. Using Strathprints, the University of Strathclyde repository as a case study, this submission will briefly review the techniques and technical changes deployed on Strathprints and examine the impact of these changes by studying data on web impact, COUNTER usage and web traffic over a 4-year period. Analysis of this unique dataset provides persuasive evidence that specific enhancements to the technical configuration of a repository can generate substantial improvements in its content discovery potential and ergo its content usage, especially over several years. In this case study COUNTER usage grew by 62%. Increases in Google 'impressions' (266%) and 'clicks' (104%) were a notable finding too, with high levels of statistical significance found in the correlation between clicks and usage ($t = 14.30$, $df = 11$, $p < 0.0005$). Web traffic to Strathprints from Google and Google Scholar was found to increase significantly with growth on some metrics exceeding 1300%. Although some of these results warrant further research, the paper nevertheless demonstrates the link between repository optimization and the need for open repositories to assume a proactive development path, especially one that prioritises web impact and discovery.

Keywords: open repositories, resource discovery, Open Access, content visibility, repository optimization, information retrieval

1 Introduction

The theme and sub-themes of OR2019 highlight the necessity of repositories to fulfil user needs and expectations. More than ever before users expect to discover open content easily, normally via search, and for their own content (typically scholarly content deposited in an open repository) to be equally discoverable. Repositories are—and have been—well placed to meet these needs but cannot remain static, isolated systems, removed from the changing technical expectations of discovery tools. This submission contributes to the discussion surrounding user discovery needs and provides evidence of the need for repository managers and developers to prioritize discovery.

Better meeting user expectations is crucial to preserving the relevance of repositories as nodes within open science infrastructure. The emergence of proprietary scholarly communications platforms represents a significant future challenge for open repositories. Such platforms are increasingly demonstrating popularity

within research institutions yet can simultaneously demonstrate poor support for open standards or prevalent open science technical protocols. Low levels of integration with existing open scholarly infrastructure is also recognised to be a frequent challenge (de Castro, 2017). Ensuring that repositories can continue to expose content as optimally as possible to search and discovery agents—and in a manner superior to alternative platforms—is a key tenet of repositories and central to their relevance to users. Understanding the way in which this can be technically achieved is important; COAR's conceptions of Next Generation Repositories (COAR, 2017) has delivered an important development path for repositories to follow in coming years. But the need to gather evolving evidence remains a necessity to direct new or unexpected streams of technical work and steer institutional decision making in instances where HEIs are confronted with decisions about selecting or migrating scholarly communications platforms.

Using Strathprints, the University of Strathclyde repository as a case study, this paper examines data on web impact, COUNTER usage and web traffic over a 4-year time-frame. The data presented were captured following the embedding of several technical adjustments and improvements to Strathprints, which have been documented in more detail in previous work (Macgregor, 2019) but which will nevertheless be briefly summarised in section 3, along with some related work in section 2. Data are described in section 4 as is its collection and analysis. However, the principal contribution of this paper, described in sections 4 and 5, is to report on the insights this longitudinal dataset provides about repository visibility and discoverability, and to deliver robust conclusions which can inform similar strategies at other institutions.

2 Repository visibility & discovery : a brief background

Previous work has noted the importance of repositories in promoting open scholarly communication and the discovery of open research content, e.g. (Arlitsch and OBrien, 2012; Tonkin et al., 2013; Kelly and Nixon, 2013; Pekala, 2018). Arlitsch (2017), in particular, provides a useful contribution on the role of search engine optimization (SEO), the importance of 'white hat' adjustments and its role in promoting repository indexing by common search engines, as well as academically focused discovery tools like Google Scholar. Contributions have also come from the individuals closer to the systems which refer much of the web traffic repositories seek. Acharya (2015), for example, delivers recommendations on repository optimization from a position of authority, noting how common technical failings inhibit satisfactory Google Scholar crawling and indexing. Yet despite these contributions to the literature—and despite the importance of repositories and their infrastructure in exposing open research content—wider understanding about repository visibility and discoverability remains embryonic. Few studies have sought to codify and then evaluate the impact of their approaches.

Recent related work by the present author has gone some way to addressing this by studying and codifying specific technical adjustments and improvements which can be made to an open repository, followed by the observation of longitudinal web and usage data in order to assess their efficacy (Macgregor, 2019). Preliminary experiments documented in Macgregor (2017) noted some encouraging evidence about the positive impact of certain repository enhancements but the small nature of the study and dataset provided only indicative results. Results from a subsequent and more detailed study from the same stream of work (Macgregor, 2019) concluded that web traffic, search traffic and COUNTER usage could be improved on the most important search and discovery tools by deploying the specified technical changes. Strong correlations between Google search visibility and repository COUNTER usage were demonstrated, as were significant increases in web traffic, Google 'impressions' and 'clicks' and COUNTER usage.

This brief paper seeks to continue the aforementioned line of enquiry by validating the results reported in Macgregor (2019) through examination of a larger web impact and COUNTER usage dataset. Analyses performed on such a large dataset better delivers reliable and actionable conclusions which can then inform repository discovery strategies elsewhere.

Table 1: Summary of technical 'adjustments' and 'improvements' implemented on Strathprints. Full details in (Macgregor, 2019)

Key technical 'adjustments'
Modification of file-naming conventions
'Minification' of all relevant repository source files
Rationalisation of all CSS and Javascript (JS) files in order to remove unused rules and variables
Asynchronous loading of JS resources
Deployment of GZIP compression
Image optimization, e.g. compression, use of .webp, etc.
Migration to InnoDB as the MySQL storage engine
Deployment of Google Data Highlighter
Key technical 'improvements'
Repository user interface (UI) improvements
'Mobile first', responsive re-engineering of repository to align with new weighting in PageRank, etc.
'White hat' improvements, e.g. navigation, hyperlink labels, content improvements promoting user interaction
'Connector-lite' ecosystem implemented within repository-CRIS interactions

3 Adjusting & improving case study : Strathprints

The case study repository for this paper, Strathprints[1] - the University of Strathclyde institutional repository, is powered by EPrints (version 3.3.13). Though EPrints is the focus here, it is thought that most of the adopted technical changes are equally applicable to other repository platforms.

Prominent repository platforms (e.g. EPrints, DSpace, Digital Commons, OJS, etc.) continue to demonstrate out-of-the-box support for discovery and interoperability with key academic tools, e.g. Google Scholar, scholarly aggregators like CORE and BASE, etc. However, there nevertheless remains wide variation in the relative visibility and discoverability of repository content, even across similar or the same repository platforms, such that it is necessary to take steps towards repository optimization. To effect change in web visibility and user engagement, thereby improving usage, a series of technical 'improvements' and 'adjustments' were implemented on Strathprints in March 2016.

'Improvements' were changes that resulted in substantive modifications to repository functionality, while 'adjustments' included actions that sought to refine existing aspects of the repository. As this paper is largely concerned with the effect of the technical changes and the resulting data, the nature of the adjustments and improvements are only summarised in Table 1 to provide context. Full details, including the motivation behind these changes, are instead available from Macgregor (2019).

4 Web impact data & results

A variety of metrics were monitored in order to measure the influence of the technical 'adjustments' and 'improvements' to Strathprints, including search traffic data from Google Search Console[2], COUNTER compliant usage data from IRUS-UK[3], Google Analytics[4] tracking data and routine statistical data from Strathprints itself. Data were captured for the year up to March 2016, representing Year 1 (Y1 = 2015/2016). This ensured a data baseline for repository web impact prior to the implementation of the technical changes. Data were then monitored for the same periods during Year 2 (Y2 = 2016/2017), Year 3 (Y2 = 2017/2018) and Year 4 (Y4 = 2018/2019), with data collection ending on 31 March 2019. It should be noted alternative temporal segmentations were used on this occasion thereby controlling for data variations potentially resulting from semester cycles, vacations, and so forth. For example, in this instance the year up to March 2016 is examined, and the same period in each subsequent year. Related prior work instead analysed data

[1]Strathprints: https//:strathprints.strath.ac.uk/
[2]Google Search Console: https://www.google.com/webmasters/tools/home
[3]IRUS-UK: https://irus.jisc.ac.uk/
[4]Google Analytics: https://analytics.google.com/

Table 2: Data table of total and unique web traffic to Strathprints during Y1-Y4, alongside total and unique traffic referred via Google and Google Scholar (GS).

	Total	Unique	Google	Unique Google	GS	Unique GS
Y1	296,200	226,791	17,436	13,274	6,208	4,827
Y2	365,024	276,042	164,550	130,565	72,179	55,294
Y3	450,520	346,851	230,953	182,227	104,051	80,786
Y4	489,140	383,117	274,983	217,826	125,405	94,305
Total Y1-Y4	1,600,884	1,232,801	687,922	543,892	307,843	235,212
% growth (Y2)	23.24	21.72	843.74	883.61	1062.68	1045.51
% growth (Y3)	23.42	25.65	40.35	39.57	44.16	46.1
% growth (Y4)	8.57	10.46	19.06	19.54	20.52	16.73
% growth (Exc. Y1)	34	38.79	73.74	70.55	67.11	66.83
Total % growth (Y1-Y4)	65.14	68.93	1477.1	1541	1920.05	1853.7

Table 3: Measures of central tendency for total and unique web traffic to Strathprints during Y1-Y4 ('Current data - A'), alongside total and unique traffic referred via Google and Google Scholar (GS). Data also include measures for 'Prior data - B' using data reported in Macgregor (2019) for comparison. Bottom row, 'Current data - A*', are 'Current data - A' data excluding outlying Y1 data.

Current data - A	**Total**	**Unique**	**GS**	**Unique GS**	**Google**	**Unique Google**
Mean (M)	400,221	308,200.3	76,960.75	58,803	171,980.5	135,973
Standard deviation (SD)	86,594.41	70,161.76	51,992.13	39,451.94	112,585.5	89,300.31
Prior data - B	**Total**	**Unique**	**GS**	**Unique GS**	**Google**	**Unique Google**
Mean (M)	386,908	296,311	83,569.33	63,691.33	196,783.67	154,834.67
Standard deviation (SD)	95,203.59	73,250.7	27,735.22	22,046.71	50,429.38	38,672.46
Current data - A*	**Total**	**Unique**	**GS**	**Unique GS**	**Google**	**Unique Google**
Mean (M)	434,894.67	335,336.67	100,545	76,795	223,495.33	176,872.67
Standard deviation (SD)	63,516.21	54,458.23	26,785.65	19,809.36	55,592.94	43,876.21

based on a typical academic calendar year (years up to end July) (Macgregor, 2019) and years up to end June (Macgregor, 2017).

4.1 Web traffic

Measurement of web traffic and unique web traffic was performed using Google Analytics (GA). Data are set out in Table 2.

Traffic in Y2 increased by 68,824 to 365,024, equating to a 23% improvement when compared to Y1. A 22% improvement in unique traffic was also observed ($n = 276,042$). Y3 also yielded a 23% increase in traffic on Y2 ($n = 450,520$), with percentage growth in unique traffic equivalent to 26% ($n = 346,851$). The increase in traffic and unique traffic for Y4 was lower than Y3 at 9% and 10% respectively.

These increases in traffic initially appear to be lower than those reported previously in Macgregor (2019) which, for example, reported a Y2 traffic increase of 54%, from 150,408 to 428,407, considerably higher than the 23% improvement reported here. Similar disparities can be observed for Y3 data too. However, it should be noted that the alternative segmentation of annual web impact data have altered the spread of traffic data across years, making direct comparisons to previous results problematic. Indeed, while Macgregor (2019) reported a plateauing of traffic (6%) and unique traffic (8%) in Y3, this paper instead reports a considerable percentage increase at 23% and 26% for Y3, with plateauing of traffic (9%) and unique traffic (11%) observed in Y4. This means that total percentage growth during the entire reporting period of this present study was more significant, at 65% and 69% for traffic and unique traffic respectively. This actually exceeds previously reported results but highlights the differences which are arise from studying different 'annual segments' of data.

Its dominance in search is such that Google is frequently found to be at the centre of many users'

information seeking strategies (Ian Rowlands et al., 2008). The results from this study do not appear to challenge this continuing assertion, nor results reported in Macgregor (2019), as Google was once again found to be the single largest referral source during the reporting period, accounting for 56% of all repository traffic in Y4. Over the entire reporting period this referral traffic (including unique traffic) increased by circa 1500% (Table 2). The most significant referral source thereafter was found to be Google Scholar (GS), equivalent to 26% of all web traffic by Y4 and growing by 1920% during the entire reporting period (Table 2). Much of this massive percentage growth can be observed in Y2, owing to a low baseline in GS traffic during Y1 but with significant increases observed in Y3 and Y4 also.

To verify the influence of outlying data points it is worthwhile briefly reviewing the extent of data variability using some common measures of central tendency. Table 3 sets out measures[5] for the total traffic data detailed above in Table 2 ('Current data - A') alongside the same measures for data reported in previous work (Macgregor, 2019), labelled in Table 3 as 'Prior data - B'. Data used for 'Prior data - B' are publicly available (Macgregor, 2018).

A higher mean and lower standard deviation for total (M_A = 400,221; SD_A = 86,594. M_B = 386,908; SD_B = 95,203) and unique traffic (M_A = 308,200; SD_A = 70,162. M_B = 296,311; SD_B = 73,251) can initially be observed within 'Current data (A). When Google and GS are considered separately, however, we notice the opposite, with lower mean traffic and higher levels of variability around the mean, highlighting the low baselines in Y1 for both Google and GS.

By excluding Y1's outlying data from these measures–as we have done in the bottom row of Table 3–we can note a higher mean, and less variability around the mean, for total (M_* = 434,895; SD_* = 63,516) and unique traffic (M_* = 335,337; SD_* = 54,458). Similarly, higher means and lower deviations for Strathprints traffic and unique traffic from Google Scholar can be observed. Interestingly, while higher means are observable for traffic and unique traffic from Google, a slighly higher standard deviation is found when compared to 'Prior data - B'.

It is significant to note from Table 2 that the traffic gains to Strathprints from GS during the reporting period experienced a more rapid rate of growth when compared to the general population of other web traffic sources. Even if we were to consider the large growth observed in Y1-Y2 as anomalous and exclude it from data as an outlier, a 74% and 70% increase in GS referral traffic and unique traffic respectively can nevertheless be observed between Y2 and Y4. This exceeds the growth rates in total (34%) and unique total traffic (39%) by some margin. Rapid growth in referral traffic from Google itself can also be found to have increased by 67% and 69% for traffic and unique traffic respectively. This is clearly lower than the figures for GS but nevertheless exceeds the growth rates observed in the wider pool of referral sources and may explain the higher standard deviation noted in 'Current data - A*'. The especially steep increase in GS traffic and unique traffic can perhaps best be observed by the profile of the chart presented in Figure 1.

4.2 Repository content discovery & usage

Search metrics offer an appropriate measure of repository content discoverability. Google Search Console was therefore used to capture search data during the reporting period, thereby enabling the effect of the technical adjustments and improvements to be explored on Google search queries. The distinction between 'impressions' and 'clicks' is recognised by Search Console and is reflected in its search data. Impressions are stated as arising when "A link to a URL record ... appears in a search result for a user", while a click is "any click that sends the user to a page outside of Google Search" (Google, 2019).

Improvements in impressions and clicks were observed in Y2 at 16% ($n = 4,537,744$) and 23% ($n = 153,539$) respectively when compared to the Y1 period. This upwards trend accelerated in subsequent reporting years. In Y3 a 69% ($n = 7,687,550$) and 21% ($n = 185,232$) increase in impressions and clicks respectively can be observed, followed by an 86% ($n = 14,290,059$) and 61% ($n = 298,020$) increase in Y4. This general upwards trend in impressions and clicks, including the aforementioned acceleration in Y3 and Y4, can be observed in Figure 2.

[5]Interquartile range has been omitted owing to the small number of cases.

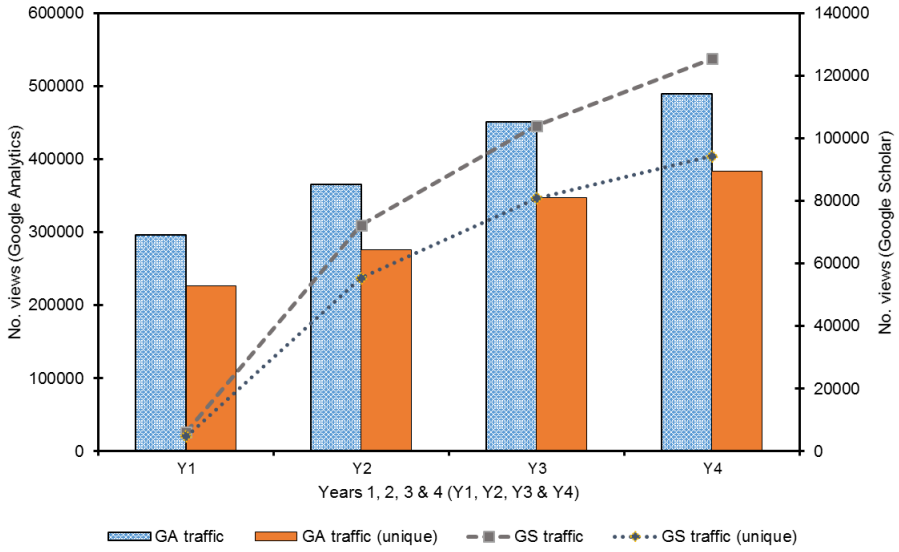

Figure 1: Volume of Google and Google Scholar referral traffic , including unique traffic in Y1, Y2, Y3 & Y4.

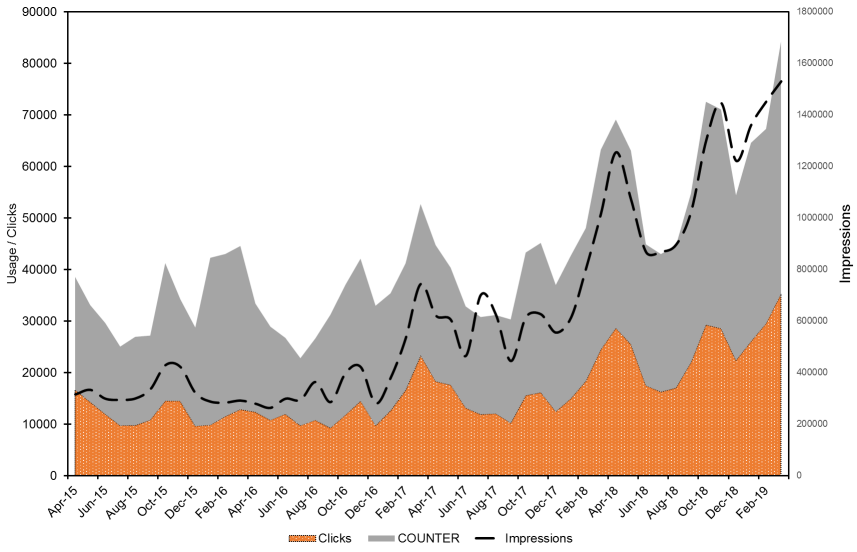

Figure 2: Strathprints COUNTER usage during Y1-Y4 alongside Google clicks and impressions during the same period.

Data are contained in Table 3. The total percentage growth in impressions and clicks during the entire reporting period was 266% and 104% respectively. Figure 3 summarises the increase in clicks, impressions and COUNTER usage - sharper increases in impressions and clicks can be noted between Y2 and Y4.

Strathprints demonstrated a 62% growth in COUNTER compliant usage during the full period examined (i.e. Y1-Y4). It is noteworthy that this growth was observed despite only a 23% growth in full-text deposits during the same period. Even where embargoed content is factored into total full-text deposits, growth remained lower (54%) than the overall increase in usage. As noted in previous work (Macgregor, 2019), usage appears to demonstrate a more nuanced pattern when it is examined on a year by year basis. Usage in Y1-Y2 is particularly notable since it deviates considerably from the results reported previously and

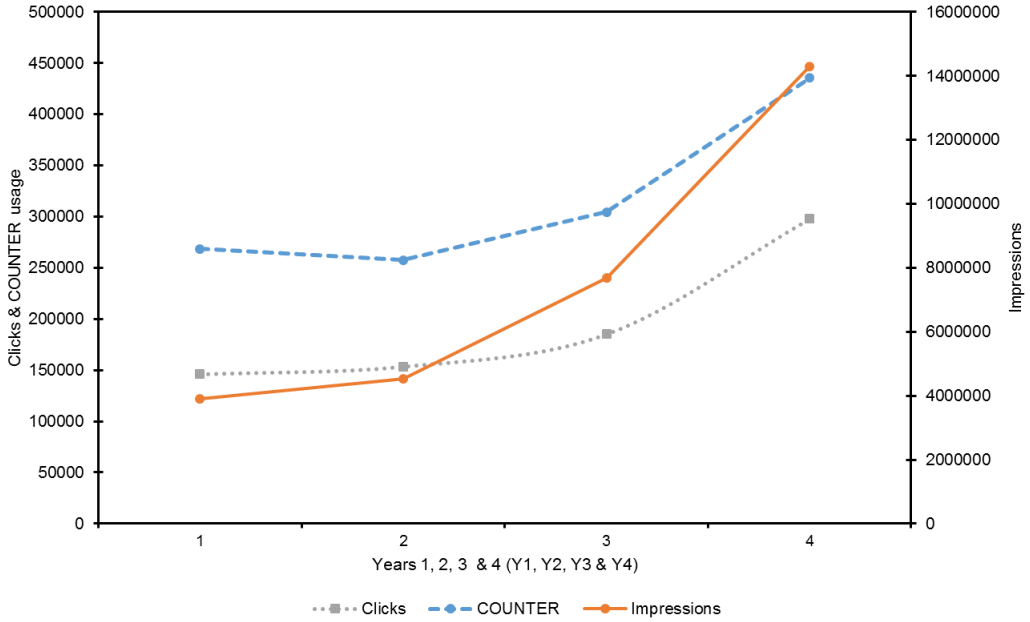

Figure 3: Charted data on observed clicks, impressions and COUNTER usage during Y1, Y2, Y3 & Y4.

Table 4: Data table of Strathprints COUNTER usage during and Google clicks and impressions during Y1-Y4. Volume of full-text OA deposits and volume of combined full-text and embargoed deposits.

	Impressions	Clicks	Usage	Deposits (OA)	Deposits (OA & Emb.)
Sub-total (Y1)	3,903,830	146,064	268,453	2,326	2,346
Sub-total (Y2)	4,537,744	153,539	257,560	2,978	3,074
Sub-total (Y3)	7,687,550	185,232	304,327	2,314	3,010
Sub-total (Y4)	14,290,059	298,020	435,467	2,861	3,620
Total (Y1-Y4)	**30,419,183**	**782,855**	**1,265,807**	**10,479**	**12,050**
% growth (Y2)	16.24	5.12	-4.06	28.03	31.03
% growth (Y3)	69.41	20.64	18.16	-22.3	-2.08
% growth (Y4)	85.89	60.89	43.09	23.64	20.27
Total % (Y1-Y4)	**266.05**	**104.03**	**62.21**	**23**	**54.31**

indicates that in the first year of observation Strathprints actually demonstrated negative growth, albeit minor. Conversely, Y4 yielded a 43% increase in COUNTER usage with only a 20% increase in full-text deposits recorded. Similarly, Y3 yielded an 18% increase in usage but experienced negative growth in full-text deposits (-22%).

It necessary to state that the cumulative effect of a mounting corpus of full-text content (with full-text deposits accumulating year upon year) is not necessarily observable in a single year of observation. It is highly probable that content deposited in Y2 benefited usage metrics in subsequent years since factors critical in discovery and usage (e.g. search engine indexing, content aggregation, etc.) can take many months. Total percentage growth across all years (i.e. 62%) is therefore a more reliable indicator of the underlying pattern. It is also apposite to highlight data from the previous section that Google search referrals and GS traffic increased well in excess of the full-text deposit rate, at 266% and 104% respectively; ergo the percentage of users being referred increased at a higher rate than the rate of full-text deposit during the reporting period. This is relevant because, based on these observations, it suggests that the rapid growth in search referrals from Google and GS has been a key factor influencing the increase in COUNTER usage.

To determine whether a correlation between Google clicks and COUNTER usage was present, Pearson's

correlation coefficient was calculated for each year in the reporting period. A correlation was detected, ranging from a weak relationship in Y1 ($r = 0.11$) to a moderate positive correlation in Y2 ($r = 0.65$). Y1 and Y2 were followed by a strengthening of the relationship in Y3 ($r = 0.87$) and Y4 ($r = 0.97$). This strengthening of the positive correlation was confirmed via the t statistic for both Y3 ($t = 5.72, df = 11, p < 0.0005$) and Y4, at a far higher level of statistical significance ($t = 14.30, df = 11, p < 0.0005$).

Computing the coefficient of determination (r^2) allows for better appreciation of the proportion of variance observed in the dependent variable (i.e. COUNTER usage) which is then predictable from the independent variable (i.e. Google clicks). In computing the coefficient of determination it was found that r^2 was significantly stronger in Y2 ($r^2 = 0.423$) than Y1 ($r^2 = 0.012$), but at such a low level that only 42% of variance in usage could be attributed to clicks. Variance narrowed considerably for Y3 ($r^2 = 0.766$) with a strong linear relationship between variables noted. This variance then narrowed again in Y4 ($r^2 = 0.953$), whereupon 95% of usage could be attributed to Google clicks. The incremental narrowing in variation between Y1 and Y4 can easily be observed from Figure 4, in which data points in Y3, and particularly Y4, are grouped more closely to the regression line.

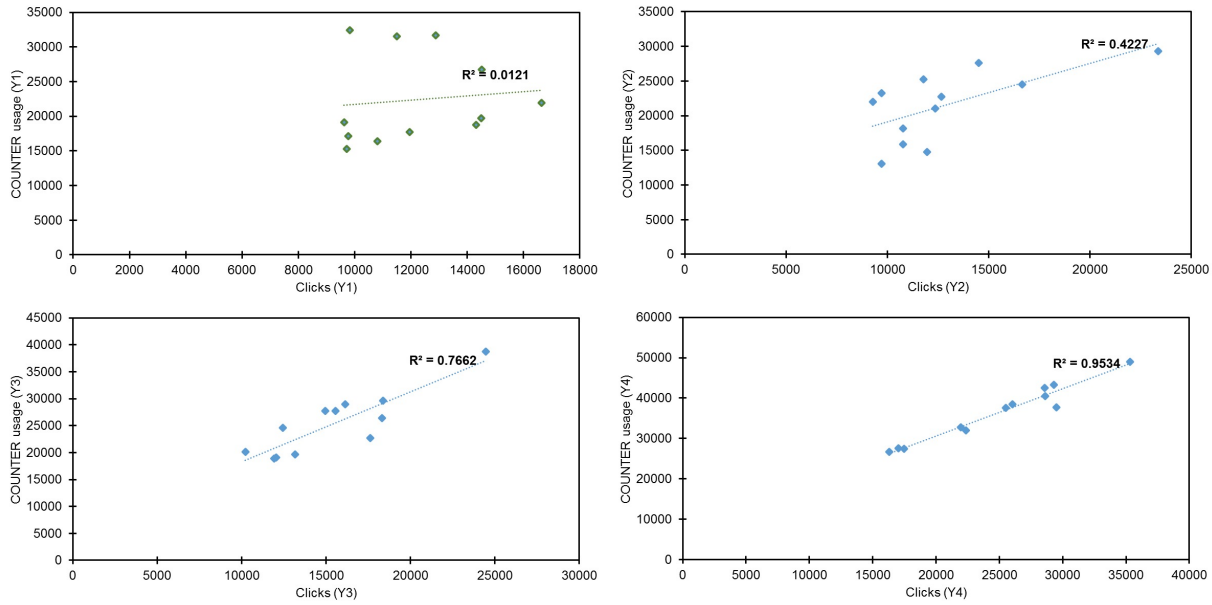

Figure 4: Coefficient of determination (r squared) for Y1, Y2, Y3 & Y4 between clicks and COUNTER usage.

5 Discussion & conclusions

This brief paper provides further analysis of the influence repository optimization approaches can have on the relative visibility, discovery and usage of an open repository. The nature of the longitudinal dataset used to track web traffic, usage and search metrics can be said to add additional weight to our findings and analysis. It corroborates previous evaluative studies (Macgregor, 2019) and reinforces prior evidence that specific technical enhancements to a repository can yield significant gains in web impact and usage.

Total web traffic was found to have increased by 65% during the period examined, with unique traffic growing 69%. Within this total and unique traffic from Google increased in excess of 70% during the reporting period, even where outlying data in Y1 were removed. Again with Y1 excluded, 67% increases in total and unique traffic were noted for Google Scholar (GS). All of this was noted despite far lower rates of full-text deposit during the reporting period. Temporal variations in when data were collected were noted as influencing some of the results and suggests that future work, or replicative studies, should attempt analyses

over a different annual reporting lifecycles.

Y1 data were excluded from some of the web traffic analyses in section 4.1 owing to their assumed anomalous appearance within subsequent data and underlying trending. It is worth revisiting this assumption here as the low baseline traffic detailed in Table 2 may have been outlying but not anomalous. Given the issues some repositories experience in achieving deep indexing by GS (e.g. Arlitsch and OBrien (2012), Acharya (2015)), and the low indexing recorded by some repositories in the recent Ranking Web of Repositories of May 2019 (CSIC, 2019), it appears quite conceivable that the low traffic baseline for Strathprints was an accurate reflection of the GS indexing penetration of Strathprints prior to the technical changes in 2016. If this were the case then percentage increases of 1920% and 1854% in total and unique traffic respectively on GS were achieved during the reporting period, attributable to the technical improvements deployed, and reflect the rapid deep indexing of Strathprints by GS. It is relevant to highlight this since it suggests that significant growth in traffic from GS is possible if steps are taken to optimize accordingly. Such high levels of indexing do also appear to be corroborated by recently published data in which Strathprints was placed in the top 5% of UK repositories and the top 10% of world repositories for number of records indexed by GS (CSIC, 2019).

But while traffic originating from GS grew considerably–and GS indexing penetration also appears to be high–it is evident that the proportion of traffic originating from GS may actually be lower than those reported elsewhere. For example, OBrien et al. (2016), who previously examined the web traffic received by four repositories, found 48%-66% of traffic to be referred by GS, which is far greater than the 26% reported in this current study. Possible explanations for this GS traffic disparity could be positive rather than negative. For instance, it is conceivable that the technical strategies deployed on Strathprints were unusually successful in promoting traffic from competing search and discovery tools such that the proportion of GS traffic appears smaller than it otherwise might. In other words, it is less that traffic from GS is less than it should be and more that the changes implemented have yielded a far greater improvement in search tools relative to GS. This would correspond with prior observations (Macgregor, 2017). Web traffic from Google certainly increased at a faster rate than GS; but it should be noted that it also started from a higher baseline in Y1. Another possible cause could be latency in detecting traffic resulting from the improved indexing of Strathprints by GS. This explanation posits that GS traffic will increase in forthcoming months and years as improvements in indexing depth and coverage translate into greater numbers of GS users being referred to Strathprints content over time. This hypothesis is something that can be easily verified by the present author and is a metric which will be monitored in future work, including any replicative studies.

A 62% increase in COUNTER compliant usage was reported despite far lower rates of full-text deposit, and even a decline in deposits during Y3. The rapid growth in search referrals from Google and GS was noted as a key driver in the overall increase in COUNTER usage during the reporting period as was their share of the total traffic Strathprints receives. This too was reflected in Google specific search metrics in which increases of 266% and 104% were observed in Google impressions and clicks respectively. The influence of Google clicks on COUNTER usage was verified via Pearson's correlation coefficient. This noted a strengthening of the relationship in every year, with high levels of statistical significance noted in years 3 and 4 (e.g. $p < 0.0005$) and r^2 demonstrating a strong linear relationship by Y4. However, the finding from this analysis that circa 95% of usage could be attributed to Google clicks warrants further scrutiny since it appears to demonstrate a potential disconnect with web traffic figures. Certainly a strong correlation exists - and this should provide a strong steer in how repositories should be developed technically over coming years. The reported growth of Google and GS traffic clearly exceeded other traffic sources, and the increase in impressions and clicks was also significant. 56% of all web traffic may have arrived via Google but the predictive potential of this analysis seems slightly incongruous ($r^2 = 0.953$), suggesting that further data gathering or replication, preferably using different repositories, could be beneficial in verifying this finding.

There are of course limitations in the way this evaluation was approached and in the data collected. Experiments seeking to effect change on third party systems are immediately problematic since it becomes impossible to control for all variables hypothesised to influence web visibility. It is therefore not claimed that every known variable has been controlled in the work for this brief paper; however, through exhaustive prior work Macgregor (2019), efforts have been taken to control as much as possible for all known variables.

Although it has been noted that Google accounted for the largest proportion of search traffic, the use of Google Search Console as a source of search metric data also presents a data compromise by excluding metrics from other discovery tools. This decision was necessary owing to the lack of data available from other discovery tools and could therefore be described as a necessary limitation. It is perhaps also worth noting that the brief nature of this conference contribution precludes any additional data analysis; additional analyses were conducted but are not presented here owing to space limitations. Interested readers are nevertheless encouraged to download the raw data for analysis and potential new insights.

Despite the limitations and some of the questions surrounding the findings, this paper provides persuasive evidence that open repositories should be managed in such a way as to enable routine technical enhancements to be deployed frequently and in response to intelligence on search, usage and web impact data. As noted in section 1, repositories cannot remain static nodes in open scholarly communications infrastructure but instead active and responsive, driving content discovery, and usage and thereby better satisfying users' needs, while simultaneously addressing the challenges presented by proprietary systems.

Data statement - Data underpinning this work are available under a Creative Commons Attribution (CC-BY) license at: https://doi.org/10.5281/zenodo.3146554

Please cite as: Macgregor, G. (2019). *Data from Promoting content discovery of open repositories : reviewing the impact of optimization techniques (2016-2019).* (pp. 1-10). Glasgow: University of Strathclyde [Strathprints repository]. Available: https://doi.org/10.17868/67963

References

Anurag Acharya. 2015. *Indexing repositories: pitfalls and best practices.* Indiana University. https://media.dlib.indiana.edu/media_objects/9z903008w

Kenning Arlitsch. 2017. Driving Traffic to Institutional Repositories: How Search Engine Optimization can Increase the Number of Downloads from IR. https://doi.org/10.5281/zenodo.894564

Kenning Arlitsch and Patrick OBrien. 2012. Invisible institutional repositories: addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech* 30, 1 (March 2012), 60–81. https://doi.org/10.1108/07378831211213210

COAR. 2017. *Next Generation Repositories: Behaviours and Technical Recommendations of the COAR Next Generation Repositories Working Group.* Technical Report. COAR, Gttingen. https://www.coar-repositories.org/files/NGR-Final-Formatted-Report-cc.pdf

CSIC. 2019. TRANSPARENT RANKING: Institutional Repositories by Google Scholar (May 2019) | Ranking Web of Repositories. https://repositories.webometrics.info/en/institutional

Pablo de Castro. 2017. 7 things you should know about Institutional Repositories, CRIS Systems, and their Interoperability. https://perma.cc/69A4-TSL8

Google. 2019. *Google Search Console.* https://www.google.com/webmasters/tools/home

Ian Rowlands, David Nicholas, Peter Williams, Paul Huntington, Maggie Fieldhouse, Barrie Gunter, Richard Withey, Hamid R. Jamali, Tom Dobrowolski, and Carol Tenopir. 2008. The Google generation: the information behaviour of the researcher of the future. *Aslib Proceedings* 60, 4 (July 2008), 290–310. https://doi.org/10.1108/00012530810887953

Brian Kelly and William Nixon. 2013. SEO analysis of institutional repositories: Whats the back story?. In *Open Repositories 2013.* University of Bath. http://opus.bath.ac.uk/35871/

George Macgregor. 2017. Reviewing repository discoverability : approaches to improving repository visibility and web impact. In *Repository Fringe 2017.* University of Edinburgh / University of Strathclyde, John McIntyre Conference Centre, University of Edinburgh. https://strathprints.strath.ac.uk/61333/

George Macgregor. 2018. Supporting dataset for: Repository optimisation & techniques to improve discoverability and web impact : an evaluation. https://doi.org/10.5281/zenodo.1411207 type: dataset.

George Macgregor. 2019. Improving the discoverability and web impact of open repositories: techniques and evaluation. *The Code4Lib Journal* 43 (Feb. 2019). https://journal.code4lib.org/articles/14180

Patrick OBrien, Kenning Arlitsch, Leila B. Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda. 2016. Undercounting File Downloads from Institutional Repositories. *Journal of Library Administration* 56, 7 (Oct. 2016), 1–24. https://scholarworks.montana.edu/xmlui/handle/1/9943

Shayna Pekala. 2018. Microdata in the IR: A Low-Barrier Approach to Enhancing Discovery of Institutional Repository Materials in Google. *Code4Lib Journal* 39 (Feb. 2018). https://journal.code4lib.org/articles/13191

Emma L. Tonkin, Stephanie Taylor, and Gregory J. L. Tourte. 2013. Cover sheets considered harmful. *Information Services & Use* 33, 2 (Jan. 2013), 129–137. https://doi.org/10.3233/ISU-130705