# RAMP – the Repository Analytics and Metrics Portal

## A prototype web service that accurately counts item downloads from institutional repositories

Authors: Patrick OBrien, Kenning Arlitsch, Jeff Mixter, Jonathan Wheeler, Leila Sterman

OBrien, Patrick, Kenning Arlitsch, Jeff Mixter, Jonathan Wheeler, and Leila Belle Sterman. "RAMP–the Repository Analytics and Metrics Portal: A prototype web service that accurately counts item downloads from institutional repositories." Library Hi Tech 35, no. 1 (2017): 144-158. DOI 10.1108/LHT-11-2016-0122

Received 4 November 2016
Revised 4 November 2016
Accepted 26 November 2016



# RAMP – the Repository Analytics and Metrics Portal

## A prototype web service that accurately counts item downloads from institutional repositories

Patrick OBrien and Kenning Arlitsch
*Library, Montana State University, Bozeman, Montana, USA*
Jeff Mixter
*OCLC Online Computer Library Center Inc, Dublin, Ohio, USA*
Jonathan Wheeler
*Library, University of New Mexico, Albuquerque, New Mexico, USA, and*
Leila Belle Sterman
*Library, Montana State University, Bozeman, Montana, USA*

### Abstract

**Purpose** – The purpose of this paper is to present data that begin to detail the deficiencies of log file analytics reporting methods that are commonly built into institutional repository (IR) platforms. The authors propose a new method for collecting and reporting IR item download metrics. This paper introduces a web service prototype that captures activity that current analytics methods are likely to either miss or over-report.
**Design/methodology/approach** – Data were extracted from DSpace Solr logs of an IR and were cross-referenced with Google Analytics and Google Search Console data to directly compare Citable Content Downloads recorded by each method.
**Findings** – This study provides evidence that log file analytics data appear to grossly over-report due to traffic from robots that are difficult to identify and screen. The study also introduces a proof-of-concept prototype that makes the research method easily accessible to IR managers who seek accurate counts of Citable Content Downloads.
**Research limitations/implications** – The method described in this paper does not account for direct access to Citable Content Downloads that originate outside Google Search properties.
**Originality/value** – This paper proposes that IR managers adopt a new reporting framework that classifies IR page views and download activity into three categories that communicate metrics about user activity related to the research process. It also proposes that IR managers rely on a hybrid of existing Google Services to improve reporting of Citable Content Downloads and offers a prototype web service where IR managers can test results for their repositories.

**Keywords** Web analytics, Assessment, Google Analytics, Institutional repositories, Google Search Console, Log file analytics

**Paper type** Research paper

### Introduction

Institutional repositories (IR) disseminate scholarly papers in an open access environment and have become a core function of the modern research library. IR run on a variety of software platforms, with great diversity in installation, configuration, and support systems,



The authors wish to express their gratitude to the Institute of Museum and Library Services, which funded this research (Arlitsch *et al.*, 2014). The authors would also like to thank to Bruce Washburn, Consulting Software Engineer at OCLC Research, for his assistance in developing RAMP, and to Susan Borda, Digital Technologies Librarian at Montana State University, for her help with data extraction.

Library Hi Tech
Vol. 35 No. 1, 2017
pp. 144-158
Emerald Publishing Limited
0737-8831
DOI 10.1108/LHT-11-2016-0122

and many libraries attempt to track file downloads as a metric of IR success. This metric is most meaningful if the measurements are consistent and accurate, and if they measure human rather than robot traffic.

Prior research published by the authors demonstrated that "up to 58% of all human-generated IR activity goes unreported by Google Analytics" (OBrien *et al.*, 2016a, b), a service that is used by approximately 80 percent of academic libraries[1]. Google Analytics is a "page tagging" analytics service that relies on tracking code in HTML pages to register visits. The tracking code is bypassed when users are sent directly to the downloadable file (usually a PDF) in the IR, as is often the case when Google Scholar (GS) is the user's discovery service of choice. This results in significant undercounting of high-value IR file downloads.

Conversely, over-counting as a result of robot traffic can occur when "log file analytics" are utilized with open source IR platforms such as DSpace. Robots (also known as "bots") account for almost 50 percent of all internet traffic (Zeifman, 2015) and 85 percent of IR downloads (Information Power Ltd, 2013). Libraries may not have the resources needed to maintain appropriate filtering mechanisms for this overwhelming robot traffic, particularly as the bots themselves are continually changing. Libraries dependent on commercial IR platforms that utilize log file analytics must trust that the vendor has sufficient skill and resources to detect and filter robot traffic.

Several projects are being developed in the European library community to set standards and develop tools for IR statistics reporting. These include: PIRUS2, which is now funded as the IRUS-UK service (Needham and Stone, 2012); a German project called Open-Access-Statistics (Haeberli-Kaul *et al.*, 2013); Statistics on the Usage of Repositories (SURE) in the Netherlands (Verhaar, 2009); and OpenAIRE, a project of the European Union in support of open access publications, including the development of usage statistics (Rettberg and Schmidt, 2012). Some of these services provide COUNTER-compliant statistics[2] processed through their infrastructure, and make data visible that can be used for national benchmarking. As of this writing, no such service exists for North American IR.

Prior research (OBrien *et al.*, 2016b) identified and defined three types of IR downloads or views: Ancillary Pages, Item Summary Pages, and Citable Content Downloads. Only the last, Citable Content Downloads, can be considered an effective measure of IR impact since they represent file downloads of the actual articles, presentations, etc., that comprise the intellectual content of the IR. Ancillary Pages are defined as the HTML pages that users click through to navigate to the content, and Item Summary Pages are also HTML pages that contain metadata, abstracts, and the link that leads to publication file. Statistics that show views of Ancillary Pages and Item Summary Pages represent limited value in the effort to demonstrate the impact of IR on the scholarly conversation.

This research compares the log files of a DSpace IR with data compiled from Google Analytics and Google Search Console (GSC). The results show a large discrepancy between these two methods. To address the significant inaccuracies of current reporting methods, this paper introduces a prototype web service that we believe provides an accurate and simple measure of Citable Content Downloads. We call this prototype web service the Repository Analytics and Metrics Portal (RAMP). RAMP is easy to use and provides a proof-of-concept solution to acquire data that are normally difficult to access and cumbersome to maintain without considerable programming skills. Prior research confirmed that the proposed method improved Citable Content Downloads reporting by more than 800 percent for two of the four IR in the study. The other two IR study participants were unaware that 100 percent of their Citable Content Downloads were missing from their Google Analytics reporting. This "miss" amounted to 299,662 downloads in a 134-day period (OBrien *et al.*, 2016b).

## Research statement

Reports of IR activity should reflect human use. The web analytics packages built into IR software platforms rely on log file analysis and are heavily biased toward over-counting item downloads. Reasons for this include extensive access to IR content by bots, and the lack of tools necessary to identify and filter bot activity from usage reports. A prototype web service called RAMP is presented as a partial solution to the difficulty of accurately measuring IR use and impact. The RAMP prototype extracts relevant GSC data that can be combined with Google Analytics to produce accurate counts of Citable Content Downloads.

## Literature review

The amount of data that search engines must mine from the web is large and increasing, as is the number of queries they try to resolve. The indexed web is currently estimated to contain nearly five billion pages (de Kunder, 2016; van den Bosch *et al.*, 2016), and Google revealed in 2016 that it now handles "at least two trillion searches per year" (Sullivan, 2016). While it is difficult to ascertain the total size of stored data on the web, total internet traffic is an easier measure and is projected to surpass one zettabyte (1,000 exabytes) by the end of 2016 (Cisco, 2016).

Search engines could not exist without robots, also known as "crawlers" or "spiders," which constantly scour websites and retrieve information to add to search engine indices. When indexing sites, "crawlers start with a list of seed URLs and branch out by extracting URLs from the pages visited" (Zineddine, 2016). Content providers depend on these robots to help gather website content into general search engines like Google and specialty search engines like GS. Frequent crawler visits to IR are necessary for harvesting new content. The activity of these indexing robots is considered beneficial, as it is "in part a positive metric, an indication of site success" (Huntington *et al.*, 2008). While sites must allow and even encourage bots to crawl and index their pages, usage reporting of IR is only meaningful if bot activity can be filtered.

Although robots are essential to the effective functioning of search engines, not all robot traffic is well intentioned. Some robots scrape content to replicate it elsewhere, a relatively benign, if potentially unethical activity. Some are designed as malware and have entirely nefarious purposes, as evidenced by the October 2016 distributed denial of service (DDoS) attack on Dyn, an internet infrastructure company that offers Domain Name System services to resolve web addresses into IP addresses (Newman, 2016).

This research project is concerned with the sheer volume of robot traffic and the difficulty in distinguishing it from human traffic. Huntington et al. (2008) estimated that robots accounted for 40 percent of all web traffic, and by 2015 that number had risen to 50 percent (Zeifman, 2015). Worse, for those attempting to accurately report the use of IR, nearly 85 percent of all IR downloads are estimated to be triggered by robots (Information Power Ltd, 2013). One of the most difficult issues in dealing with robots is simply detecting their presence as non-human action. "Every new crawler stays unknown for a while and it is up to the detection techniques to ensure that such period is as short as possible" (Lourenço and Belo, 2006). The published research that is most aligned with ours was a two-year bot detection study that found 85 percent of the traffic to an IR was from robots (Greene, 2016). This was the first benchmark study of its kind for Open Access IR, but it focused exclusively on bot detection rather than achieving accurate counts of human download activity.

Google's share of the explicit search engine market has hovered around 65 percent (comScore Inc., 2016) for at least the past five years (comScore Inc., 2011), and the company's specialized academic search engine, GS, has become very popular among those seeking scholarly content. It is difficult to determine GS market share because "GS usage information is not available to participating institutions or libraries" (Herrera, 2011).

However, its ease of use and broad coverage has contributed to its popularity (Nicholas *et al.*, 2009) and its growth. A University of Minnesota survey of 1,141 graduate students found that over half used GS at least a few times each month (Cothran, 2011), and a San Francisco State University study found that GS was the top SFX source for requests in 2011 (Wang and Howard, 2012). A report from JISC found that "30% of doctoral students used Google or Google Scholar as their main source of research information they sought," but more specifically, the study found Google sources were "strongly favored above other sources by arts and humanities, social science and engineering and computer science students" (Carpenter, 2012). GS is also very popular among academic faculty and professional scientists. A *Nature* survey of 3,000 scholars showed that over 60 percent of scientists and engineers and over 70 percent of scholars in social sciences, arts, and humanities use GS on a regular basis (Van Noorden, 2014).

Some research shows that GS may be less reliable, updated less frequently, and has a more web traffic-based ranking than other academic indexing services, such as Scopus, Web of Science, and PubMed (Falagas *et al.*, 2008). While some studies have noted a 100 percent retrieval of sources from replicated systematic reviews (Gehanno *et al.*, 2013), others find that although the coverage in GS is quite good (~95 percent retrieval of biomedical research) it is less efficient than searching Web of Science, Scopus, or PsycINFO (Giustini and Boulos, 2013). GS is still often recommended to medical professionals for serendipitous discovery (Gehanno *et al.*, 2013).

Questions have also persisted about the parent company's commitment to the GS search engine. Although GS's Chief Engineer, Anurag Acharya, "declines to reveal usage figures, he claims that the number of users is growing worldwide, particularly in China. And the Google Scholar team is expanding, not contracting" (Bohannon, 2014). While GS has detractors, its scope and use is growing.

### Market conditions and incentive

The method and prototype service introduced in this paper leverage Google tools because they are among the best available, due to market incentives. A strategic market force that works in the library community's favor is the fact that 90 percent of Google's US$75 billion 2015 revenue was generated by its proprietary advertising network, which is based on a Pay Per Click (PPC) advertising model (Alphabet Inc., 2015). PPC relies on advertisers bidding to display ads based upon the keywords used in search queries. Every time an advertisement is clicked, the advertiser typically pays Google between US$0.05 and US$50.00, with the price determined by an efficient market facilitated by Google's real-time customer bidding system (Edelman *et al.*, 2007). The market reveals that Google's customers are unwilling to pay over $900 per click for the most expensive PPC key word (Lake, 2016) without some certainty that potential customers (not robots) are clicking their advertisement. As a result, Google is one of the best in the world at robot detection and screening. We can also assume that Google incentives to invest in bot detection will remain strong if their advertising revenue continues to grow. In short, market conditions provide Google with the incentive and resources to invest in bot detection that far exceeds the abilities of the library community.

### Tools used in this study

GSC is a free diagnostic tool that was previously known as Google Webmaster Tools; it was rebranded in 2015 to broaden its appeal and use (Google Inc., 2015). GSC captures data about queries related to websites, alerts webmasters to problems encountered by Google's crawlers, and provides a management interface for monitoring these sites. "Search Console provides actionable reports, tools, and learning resources designed to get your content on Google Search" (Google Inc., 2016a). The GSC provides various metrics for reporting,

including clicks, impressions, date, device, etc. GSC data can be accessed directly through the "Search Console" dashboard in the Google Webmaster Tools site[3], queried through the API using Python or Java, or through the Query Explorer[4]. GSC records item downloads from all Google search properties.

Apache Solr is used in DSpace to index the item-level metadata as well as the usage log data that contain page view and download statistics. Solr can be queried through the web UI using a "localhost" setup or from the command line using "curl." The Solr data contain statistics on item page-level usage as well as file-level usage. Each record is a "click" as defined by Google (Google Inc., 2016b). However, Solr data are raw data and do not provide definitions concerning timeouts, double clicking, etc. and make no attempt to tally or screen activity (i.e. downloads) per bitstream per day. The metrics Solr provides include URL, date, device, city, country, IP address, referrer, id, handle, etc. (Diggory and Lawrence, 2016).

## Research method

The research team built two data sets from different sources to allow direct comparisons of Citable Content Downloads events. The first source was a combination of Google Analytics and GSC download events, and will henceforth be referred to as the GA/GSC data set. The second data set was compiled from the Solr log files built into the DSpace platform. The GA/GSC event data include page URLs containing bitstreams that could be parsed to create an index of Citable Content Downloads for each item based on its DSpace handle. For example, the URL below has a handle of "1/9943" and is a good representation of raw data compiled from GA/GSC events (http://scholarworks.montana.edu/xmlui/bitstream/handle/1/9943/IR-Undercounting-preprint_2016-07-19.pdf?sequence=6&isAllowed=y).

Data were collected from January 5 through May 17, 2016, ($n = 134$ days) and consist of three primary sources related to the Montana State University (MSU) IR, called ScholarWorks:

(1) "Undercounting File Downloads from Institutional Repositories" data set (OBrien *et al.*, 2016a). This data set consists of daily Citable Content Downloads events ($n = 45{,}158$) collected for each URL by GSC.

(2) Citable Content Downloads events initiated though a DSpace web page and recorded by Google Analytics events. These records ($n = 5{,}640$) were extracted via the Google Analytics API and included the following dimensions and metrics:

- Event Category (ga:totalEvents).
- Page (ga:pagePath).
- Unique Events (ga:uniqueEvents).

(3) Disaggregated log data were extracted from the ScholarWorks DSpace Solr indexes (Masár, 2015). DSpace Solr indexes are divided into multiple parts consisting of a statistics core and a search core. Approximately 1.8 MM records were extracted from the ScholarWorks Solr statistics core for the 134-day research period. The search core includes metadata about communities, collections, and items. Search core data were extracted and joined with the statistics core data to associate logged Citable Content Downloads events with corresponding DSpace item handles. Listed below are the critical fields and settings required for joining the DSpace Solr statistics and search data; statistics core fields are shown in bold font and search core fields are italicized:

- **Time** = January 5, 2016-May 17, 2016
- **IsBot** = False

- **StatisticsType** = view
- **BundleName** = ORIGINAL
- **Type** = 0
- **OwningItem**: *
- **Id**: *
- *Search.ResourceType = 2*
- *Handle* = *
- *Search.ResourceId* = *
- *StreamSourceInfo* = *

These three data sets[5] were combined into a single data set ($n = 130{,}384$) representing the Citable Content Downloads events recorded by GA/GSC and the MSU DSpace IR for each handle on a daily basis.

## Limitations

The main limitation of the method described in this paper is that it does not account for Citable Content Downloads that originate outside Google Search properties. For instance, non-Google search engines (e.g. Yahoo!, Bing, Yandex, etc.) may also send users directly to the PDF file of the article in the IR, and these cases would not be recorded with the method in this study. However, Google's 65 percent market share for its combined search engine properties (comScore Inc., 2016) is quite high, and the number of direct links to Citable Content Downloads in IR from non-Google properties is therefore likely to be small. This is another area we are studying, and preliminary results have confirmed that the IR in our study receive very few direct links from Yahoo or Bing. A tool similar to GSC for Bing and Yahoo is now available, which may allow most of the remainder of the traffic commercial search engine market to be tested in a future study (Microsoft Inc., 2016).

Other Citable Content Downloads that are not included are direct links that are exchanged in e-mail or text messages, or that have been posted on web pages or on non-Google social media sites like Facebook. Again, the numbers of these links are likely to be small for a given IR, because these referrers are less likely to serve as intellectual discovery services for scholars.

Finally, the research is limited by the team's lack of resources to enhance features, such as integration with Google Analytics API data and IR metadata, and it may not be able to maintain the prototype service beyond the IMLS "Measuring Up" grant project expiration date of December 2017.

## Findings

The findings of this study demonstrate that there is an enormous disparity between the reporting methods (log file, GA, and GSC), and that they are not comparable in any way. Specifically, log file data capture all Citable Content Downloads activity in the IR and apply a standard filter used by most software vendors to screen out known bots; Google Analytics Download Events reflect all Citable Content Downloads activity that originates from Ancillary and Item Summary HTML pages internal to the IR; and GSC provides all Citable Content Downloads events that occurred via direct link from a Google search property.

The two biggest unknown factors are:

(1) How many Type I errors (false positives of non-human Citable Content Downloads events) are included in metrics generated from log data?

(2) How many Type II errors (false negatives of human Citable Content Downloads events) in IR metrics are excluded by the GA/GSC from direct links that originate from non-Google search properties?

Most notable in the descriptive statistics describing the data grouped by date (Table I) is the overwhelming disparity that shows the potential instability and lack of predictability of the activity reported by the two methods. The most concerning result is the very large Kurtosis that indicates that the log data have infrequent but large deviations typically known as outliers. The Google Kurtosis, on the other hand, is relatively close to the mean – representing a smooth distribution of activity over time. The finding that log data have a standard deviation larger than their mean and median is consistent with the Kurtosis. The data gathered through GA/GSC, on the other hand, show a relatively low standard deviation, indicating a more normal distribution of activity. The following conclusions can be drawn from Table I: data from the log files could be interpreted to claim that the MSU IR averages 380 percent more Citable Content Downloads each day than the data drawn from GA/GSC; and the standard deviation of the log file data indicates that user activity swings wildly from day to day, while the data drawn from GA/GSC are relatively stable and predictable over time.

Table II provides another look at the data aggregated by unique identifier (i.e. DSpace handle) over the entire 134-day period. Again, relevant descriptive statistics shown indicate a large difference between the two tracking methods. Because of the log data outliers, the median is more relevant than mean. However, the log data median is also significantly larger than the Google mid-point. The log data standard deviation and range are almost

**Table I.**
Descriptive statistics for Citable Content Downloads by date

| | Log data | Google |
|---|---|---|
| Mean | 2,405.72 | 626.39 |
| Standard error | 249.11 | 10.22 |
| Median | 1,835.00 | 640.00 |
| Mode | 1,728.00 | 704.00 |
| SD | 2,883.68 | 118.35 |
| Variance | 8,315,621.23 | 14,006.28 |
| Kurtosis | 26.17 | −0.99 |
| Skewness | 5.10 | −0.20 |
| Range | 19,429 | 500 |
| Minimum | 492 | 370 |
| Maximum | 19,921 | 870 |
| Sum | 322,366 | 83,936 |
| Count | 134 | 134 |

**Table II.**
Descriptive statistics by handle

| | Log data | Google |
|---|---|---|
| Mean | 38.01 | 9.90 |
| Standard error | 1.51 | 0.63 |
| Median | 14 | 2 |
| Mode | 1 | 0 |
| SD | 138.64 | 58.32 |
| Sample variance | 19,222.20 | 3,401.35 |
| Range | 4,872.00 | 2,060.00 |
| Minimum | 0 | 0 |
| Maximum | 4,872 | 2,060 |
| Sum | 32,236 | 83,936 |
| Count | 8,482 | 8,482 |

140 percent more than Google's. This indicates that the issue is probably more involved than detecting and removing a few obvious robot outliers.

Looking at the *t*-Test Pairwise Two Sample for Means for data aggregated by date (Table III) or handle (Table IV) indicates a strong rejection of the hypothesis that the two methods produce similar results. Tables III and IV also contain the *F*-test Two sample for variance results that also confirm that we can reject the hypothesis that the two methods are the same.

When graphing the data by weekday (Figure 1), one would expect to see similar-sized weekday totals and trend lines. However, the two methods are divergent with log data indicating Wednesdays produces 422 percent more Citable Content Downloads events than were tracked with GA/GSC.

## Discussion

Citable Content Downloads events reported by IR are initiated by humans seeking information. Including robot activity (Type I errors) in reported IR metrics does more harm than good for strategic and operational decision making. Library stakeholders are better served by excluding some human IR activity (Type II errors) from IR

| | Log data | Google |
|---|---|---|
| Mean | 2,405.72 | 626.39 |
| Variance | 8,315,621.23 | 14,006.28 |
| Observations | 134.00 | 134.00 |
| Pearson correlation | 0.22 | |
| Hypothesized mean difference | – | |
| df | 133.00 | |
| *t* stat | 7.20 | |
| $P(T \leqslant t)$ one-tail | 0.00 | |
| *t* critical one-tail | 1.66 | |
| $P(T \leqslant t)$ two-tail | 0.00 | |
| *t* critical two-tail | 1.98 | |
| *F* | 593.71 | |
| $P(F \leqslant f)$ one-tail | 0.00 | |
| *F* critical one-tail | 1.33 | |

**Table III.**
Date *F* and *t*-test pairwise two sample for means

| | Log data | Google |
|---|---|---|
| Mean | 38.01 | 9.90 |
| Variance | 19,222.20 | 3,401.35 |
| Observations | 8,482.00 | 8,482.00 |
| Pearson correlation | 0.83 | |
| Hypothesized mean difference | 0.00 | |
| df | 8,481.00 | |
| *t* Stat | 26.94 | |
| $P(T \leqslant t)$ one-tail | 0.00 | |
| *t* critical one-tail | 1.65 | |
| $P(T \leqslant t)$ two-tail | 0.00 | |
| *t* critical two-tail | 1.96 | |
| *F* | 5.65 | |
| $P(F \leqslant f)$ one-tail | 0.00 | |
| *F* critical one-tail | 1.04 | |

**Table IV.**
Handle *F* and *t*-test pairwise two sample for means

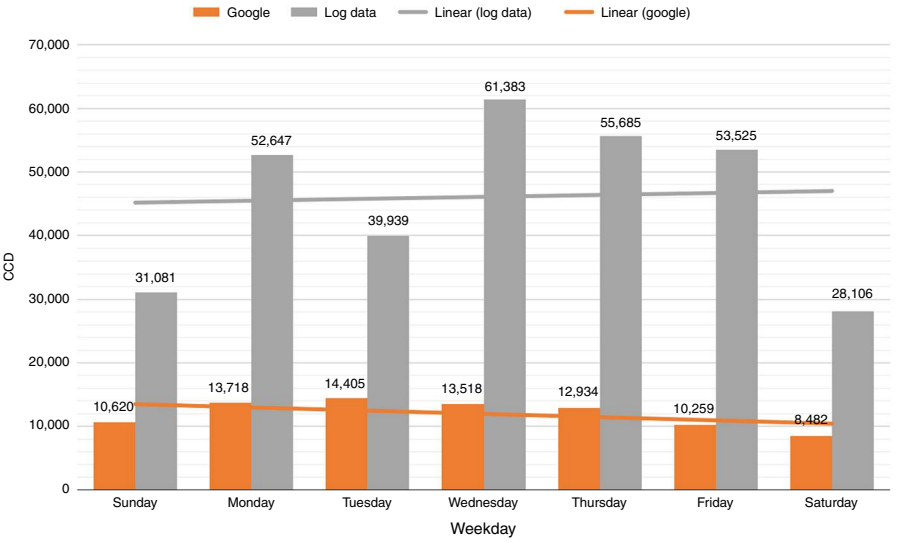


**Figure 1.**
Citable Content Downloads events by weekday

reporting metrics if those reports can also exclude bot activity (Type I errors). While Type II errors are not desirable, Type I errors give the impression of poor integrity and jeopardizes the public trust and good will the library community has with its stakeholders.

Google Analytics is the most prevalent web analytics service being used in academic libraries, and for good reason: it is powerful, free, and relatively easy to implement. However, incorporating the most important metric for IR – Citable Content Downloads – is disjointed, difficult to access and limited to a moving 90-day access window. Without the skills to systematically access the GSC API, IR managers do not have access to a persistent data set of their largest and most important metric.

We advocate the creation of a single data store similar to the IRUS-UK initiative, although our method uses web service based on GA/GSC platforms. Our research demonstrates that the proposed method has a very high potential of providing a superior long-term solution than what the library community might build and maintain on its own. There are a few caveats related to privacy and long-term access costs that require further investigation and discussion. Because our team is preparing to submit data and analysis on Google Analytics privacy for publication, we are limiting our scope of discussion in this paper to long-term access and cost.

The log file analytics packages built into IR platforms offer a solution that initially appears to capture 284 percent more Citable Content Downloads, but closer inspection reveals that much of that traffic is not generated by humans. The capacity to accurately filter robot traffic from these log file analytics packages is beyond most libraries. Worse, it is impossible to tell whether a platform that claims to effectively filter robot traffic is actually able to do so as well as Google.

## RAMP web service prototype

The RAMP prototype offers a simple web interface for IR managers to view and download Citable Content Downloads event data from their IR for a given date. The service automates the daily aggregation of Google Search API analytics and stores them as daily file dumps. It also provides single-day statistics and visualizations.

The research team reused and refactored Python code originally developed for the previously published research (OBrien *et al.*, 2016b) to experiment with the proof-of-concept prototype. RAMP provides the following capabilities:

(1) persistent access to Citable Content Downloads event data over time;

(2) no significant investment in training or system configuration; and

(3) the potential to aggregate IR metrics across organizations for consistent benchmarking and analysis interpretation.

The landing page (Figure 2) provides a list of current organizations that have registered with RAMP as of February 2017. The service provides two ways of collecting/reviewing GSC data. The first is real-time daily statistics, where users select a date and the service conducts a query of the Google Search API to retrieve data (see Figure 3). This method is limited to 90 days of data that GSC stores.

For the daily statistics, users can view the data in two different ways. The first is to download the data, which includes total clicks on Citable Content Downloads and total number of clicks per device. Downloaded data are available as TSV format and could be compiled locally for use in spreadsheets or other applications.

The second method of viewing data is through RAMP's visualization feature, which currently shows how many Citable Content Downloads URLs were clicked in a given day as well as the type of device (Mobile, Tablet, or Desktop) that accessed the document (Figure 4).

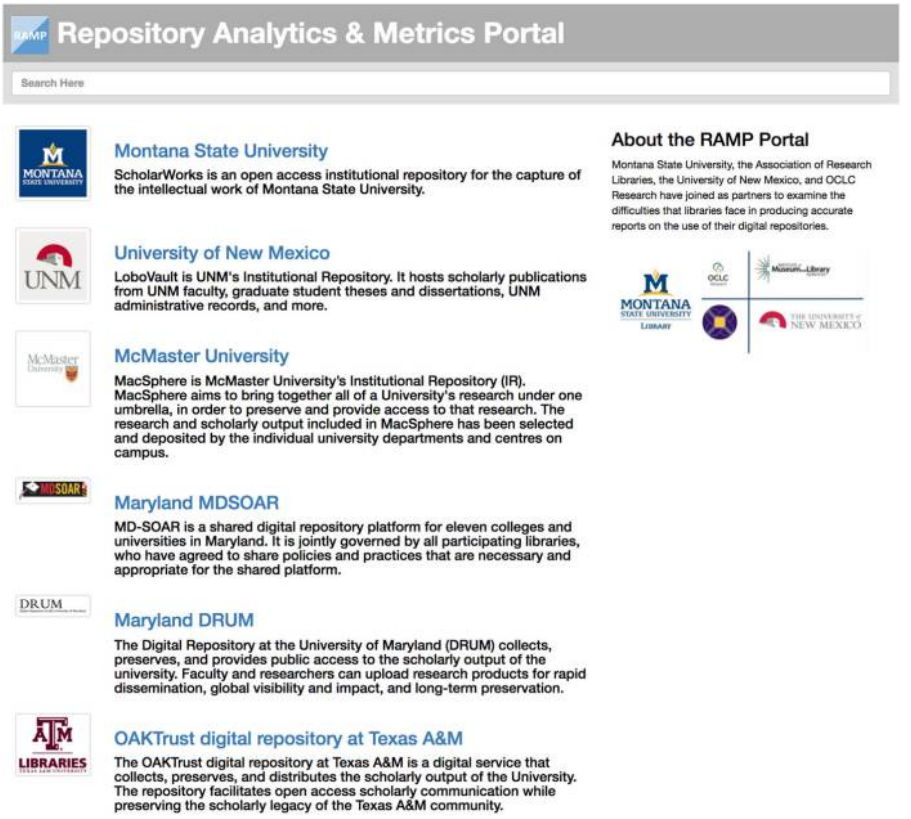


**Figure 2.**
Landing page for the RAMP service

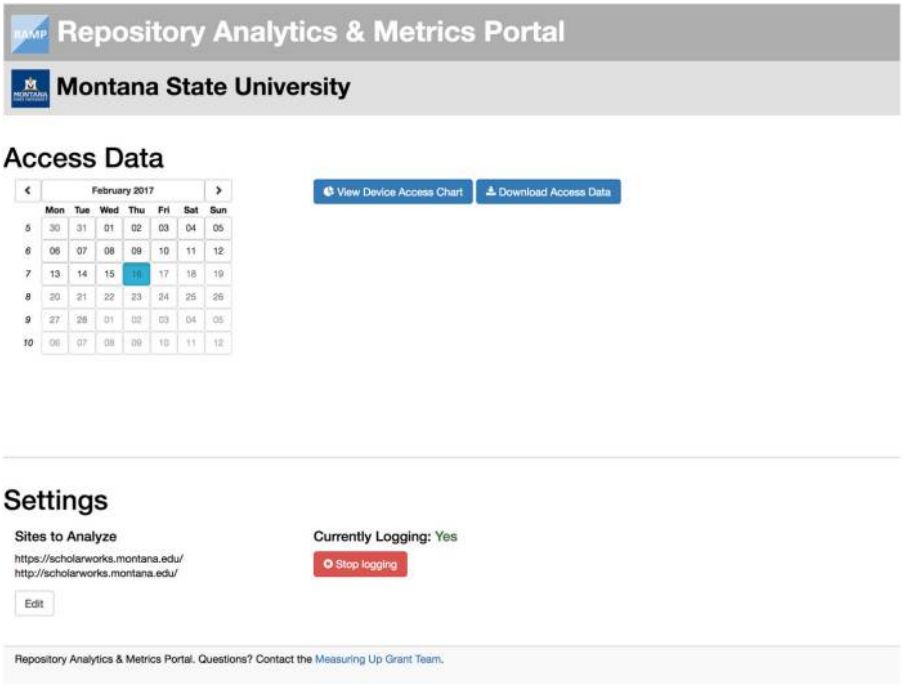


**Figure 3.**
Organization page with administrative login

In addition to querying daily statistics, we recommend users set up the service to automatically collect daily logs every day at midnight to ensure access to their data past the 90-day window imposed by Google. It should be noted that the Google Search API has a three-day delay before statistics are available for download. Consequently, the daily statistics query actually pulls data from three days prior to the current date.

## Registering for access to RAMP

The prototype currently runs on Google's Cloud Platform and leverages Google's APIs. The service requires no installations, configurations, special skills, or training, and once

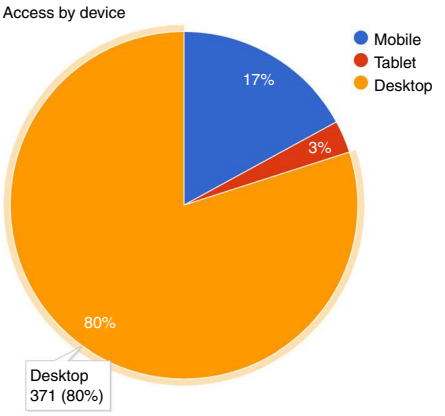


**Figure 4.**
Daily statistics visualization by device

approved, IR managers can have access to data in a matter of minutes if they already run Google Analytics or GSC. The only local requirement is that the Google Account administrator for the IR adds RAMP's e-mail to Google Analytics or GSC, as would be done for any other "user" in the organization.

IR managers should send an e-mail[6] to our research team, requesting access to RAMP. Below is the process for accessing the RAMP service.

Once authorization has been provided, RAMP developers will create a registration entry for the participating institution and forward a special RAMP service account e-mail address. To begin downloading and analyzing data, the IR manager will need to authorize RAMP to access their repository's GSC data by adding the RAMP service account e-mail address[7]. Next, they will input the repository's base URL in the "Sites to Analyze" form (Figure 5). The base URL should correspond to the web property for which the service account is authorized as described above. It is possible to add multiple URLs in order to include both HTTP and HTTPS protocols, where applicable.

Once the websites are added, the IR manager is able to view real-time daily statistics for the past 90 days as well as initiate daily logging. Because the RAMP application has no access to any personally identifiable information, proprietary, or confidential information, anyone with access to the RAMP system can search or download the daily statistics collected for any of the participating institutions. Although the system is open for all accepted participants to read, participating institutions can stop the logging of their data any time. In this pilot, the MSU will have the authority to start and stop daily logging and add or modify the URLs that are used by RAMP.

## Conclusion

Web server logs for IR platforms are excellent at tracking all activity (i.e. page views, visits, item downloads, etc.). However, analytics reports using log analysis are heavily biased toward over-reporting due to excessive utilization of IR content by robots. The problems with existing methods of reporting Citable Content Downloads from IR can be summarized into three statements: page tagging analytic services grossly undercount, locally administered log file analytics grossly over-count, and it is difficult to ascertain whether commercial services that offer log file analysis packages can manage the rapidly changing robot environment quickly enough to provide consistent measurement.

Currently, the open source repository community does not have access to a reasonable solution for identifying and filtering out this unwanted bot activity consistently. Solutions developed and tested by the community appear effective on obvious and less sophisticated robots. However, 50 percent of bot activity is due to "bad bots" (Zeifman, 2015) that require highly advanced machine learning risk assessment algorithms for real-time bot detection. These methods were pioneered and deployed by former academics who now work at Google and other leading companies (*The Economist*, 2016). Without this advanced technology the

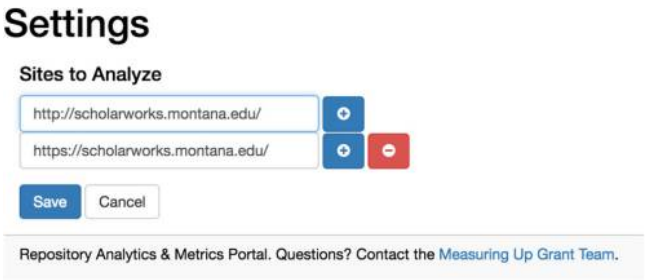


**Figure 5.**
Managing URLs used by the RAMP service

statistical sensitivity and specificity of any solutions developed by the library community is questionable, at best.

The RAMP prototype described in this paper utilizes Google services for a solution that is accessible and free. It relies on Google's platforms to ensure that all participants are using clearly defined metrics and terminology. Enhancements made to Google's platform pose little technical risk for participant implementation, and any changes in Google services affect all participants at the same time. This ensures that all participant data are comparable over time and allow the library community to aggregate data for benchmarking and best practice identification with confidence. Finally, the method provides high accuracy and robustness.

## Notes

1. Privacy research in progress by the authors shows 80 percent of academic libraries that are members of ARL, DLF, or the OCLC Research Libraries Partnership use Google Analytics. Publication expected in 2017.
2. Project counter – www.projectcounter.org/code-of-practice-sections/general-information/
3. Google Search Console dashboard – www.google.com/webmasters/tools/search-analytics
4. Google Analytics Query Explorer – https://ga-dev-tools.appspot.com/query-explorer/
5. Google Search Console + Google Analytics + DSpace Solr Cores (statistics + search)
6. Send e-mail to Jeff Mixter, Software Engineer at OCLC – mixterj@oclc.org
7. https://support.google.com/webmasters/answer/2453966?hl=en

## References

Alphabet Inc. (2015), "Consolidated revenues", Form 10K, United States Securities and Exchange Commission, Washington, DC, available at: www.sec.gov/Archives/edgar/data/1288776/000165204416000012/goog10-k2015.htm#s2A481E6E5C511C2C8AAECA5160BB1908 (accessed October 28, 2016).

Arlitsch, K., OBrien, P., Kyrillidou, M., Clark, J.A., Young, S.W.H., Mixter, J., Chao, Z., Freels-Stendel, B. and Stewart, C. (2014), "Measuring up: assessing accuracy of reported use and impact of digital repositories", Funded grant proposal, Institute of Museum and Library Services, Washington, DC, available at: http://scholarworks.montana.edu/xmlui/handle/1/8924 (accessed July 15, 2016).

Bohannon, J. (2014), "Google Scholar wins raves – but can it be trusted?", *Science Magazine*, January 3, p. 14.

Carpenter, J. (2012), "Researchers of tomorrow: the research behaviour of Generation Y doctoral students", *Information Services and Use*, Vol. 32 Nos 1-2, pp. 3-17, doi: 10.3233/ISU-2012-0637.

Cisco (2016), "The zettabyte era – trends and analysis", Cisco, Cisco Visual Networking Index, available at: www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.html

comScore Inc. (2011), "comScore releases June 2011 US search engine rankings", July 13, available at: www.comscore.com/Press_Events/Press_Releases/2011/7/comScore_Releases_June_2011_U.S._Search_Engine_Rankings (accessed August 10, 2011).

comScore Inc. (2016), "comScore releases February 2016 US desktop search engine rankings", March 16, available at: www.comscore.com/Insights/Rankings/comScore-Releases-February-2016-US-Desktop-Search-Engine-Rankings (accessed October 2, 2016).

Cothran, T. (2011), "Google scholar acceptance and use among graduate students: a quantitative study", *Library & Information Science Research*, Vol. 33 No. 4, pp. 293-301, doi: 10.1016/j.lisr.2011.02.001.

de Kunder, M (2016), "The size of the World Wide Web (the internet)", October 2, available at: www.worldwidewebsize.com

Diggory, M. and Lawrence, A. (2016), "SOLR statistics", DuraSpace, Dspace Documentation Wiki, July 11, available at: https://wiki.duraspace.org/display/DSDOC5x/SOLR+Statistics (accessed October 28, 2016).

Edelman, B., Ostrovsky, M. and Schwarz, M. (2007), "Internet advertising and the generalized second-price auction: selling billions of dollars worth of keywords", *The American Economic Review*, Vol. 97 No. 1, pp. 242-259, doi: 10.1257/000282807780323523.

Falagas, M.E., Eleni, I.P., Malietzis, G.A. and Pappas, G. (2008), "Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses", *The FASEB Journal*, Vol. 22 No. 2, pp. 338-342, doi: 10.1096/fj.07-9492LSF.

Gehanno, J.-F., Rollin, L. and Darmoni, S. (2013), "Is the coverage of Google Scholar enough to be used alone for systematic reviews", *BMC Medical Informatics and Decision Making*, Vol. 13 No. 1, doi: 10.1186/1472-6947-13-7, available at: http://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/1472-6947-13-7

Giustini, D. and Boulos, M.N.K. (2013), "Google Scholar is not enough to be used alone for systematic reviews", *Online Journal of Public Health Informatics*, Vol. 5 No. 2, pp. 1-9, doi: 10.5210/ojphi.v5i2.4623.

Google Inc. (2015), "Announcing Google Search Console – the new webmaster tools", Google Webmaster Central Blog, May 20, available at: https://webmasters.googleblog.com/2015/05/announcing-google-search-console-new.html (accessed October 29, 2016).

Google Inc. (2016a), "Using Search Console with your website", Google Search Console Help, available at: https://support.google.com/webmasters/answer/6258314?hl=en&ref_topic=3309469 (accessed October 28, 2016).

Google Inc. (2016b), "What are impressions, position, and clicks? – Search Console Help", available at: https://support.google.com/webmasters/answer/7042828#click (accessed October 28).

Greene, J. (2016), "Web robot detection in scholarly open access institutional repositories", *Library Hi Tech*, Vol. 34 No. 3, pp. 500-520, available at: http://hdl.handle.net/10197/7682

Haeberli-Kaul, J., Beucke, D., Hitzler, M., Holtz, A., Mimkes, J., Riese, W., Herb, U., Recke, M., Schmidt, B., Schulze, M., Henneberger, S. and Stemmer, B. (2013), "Standardised usage statistics for open access repositories and publication services", *DINI – Deutsche Initiative für Netzwerkinformation E.V., Göttingen* (Trans by A. Rennison), available at: http://nbn-resolving.de/urn:nbn:de:kobv:11-100217555

Herrera, G. (2011), "Google Scholar users and user behaviors: an exploratory study", *College & Research Libraries*, Vol. 72 No. 4, pp. 316-330, doi: 10.5860/crl-125rl.

Huntington, P., Nicholas, D. and Jamali, H.R. (2008), "Web robot detection in the scholarly information environment", *Journal of Information Science*, Vol. 34 No. 5, pp. 726-741, doi: 10.1177/0165551507087237.

Information Power Ltd (2013), "IRUS download data – identifying unusual usage", IRUS Download Report, available at: www.irus.mimas.ac.uk/news/IRUS_download_data_Final_report.pdf (accessed July 1, 2016).

Lake, C. (2016), "The most expensive 100 Google Adwords keywords in the US", *Search Engine Watch*, May 31, available at: https://searchenginewatch.com/2016/05/31/the-most-expensive-100-google-adwords-keywords-in-the-us/ (accessed November 2, 2016).

Lourenço, A.G. and Belo, O.O. (2006), "Catching web crawlers in the act", *Proceedings of the 6th International Conference on Web Engineering*, ACM Press, Palo Alto, CA, pp. 265-272, doi: 10.1145/1145581.1145634, available at: http://portal.acm.org/citation.cfm?doid=1145581.1145634

Masár, I. (2015), "Solr – DSpace – Duraspace wiki", Dspace Documentation Wiki, December 11, available at: https://wiki.duraspace.org/display/DSPACE/Solr#Solr-Bypassinglocalhostrestriontemporarily (accessed July 1, 2016).

Microsoft Inc. (2016), "Search keywords report", Bing Webmaster Tools, available at: www.bing.com/webmaster/help/search-keywords-report-20a352af (accessed November 3, 2016).

Needham, P. and Stone, G. (2012), "IRUS-UK: making scholarly statistics count in UK repositories", *Insights: The UKSG Journal*, Vol. 25 No. 3, pp. 262-266, doi: 10.1629/2048-7754.25.3.262.

Newman, L.H. (2016), "What we know about Friday's massive east coast internet outage", *Wired*, October 21, available at: www.wired.com/2016/10/internet-outage-ddos-dns-dyn/ (accessed October 23, 2016).

Nicholas, D., Clark, D., Rowlands, I. and Jamali, H.R. (2009), "Online use and information seeking behaviour: institutional and subject comparisons of UK researchers", *Journal of Information Science*, Vol. 35 No. 6, pp. 660-676, doi: 10.1177/0165551509338341.

OBrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J. and Borda, S. (2016a), *Data Set Supporting the Study Undercounting File Downloads from Institutional Repositories*, Montana State University, Bozeman, MT, available at: http://scholarworks.montana.edu/xmlui/handle/1/9939

OBrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J. and Borda, S. (2016b), "Undercounting file downloads from institutional repositories", *Journal of Library Administration*, Vol. 56 No. 7, pp. 854-874, doi: 10.1080/01930826.2016.1216224.

Rettberg, N. and Schmidt, B. (2012), "OpenAIRE – building a collaborative open access infrastructure for European researchers", *LIBER Quarterly*, Vol. 22 No. 3, pp. 160-175.

Sullivan, D. (2016), "Google now handles at least 2 trillion searches per year", *Search Engine Land*, May 24, available at: http://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247 (accessed October 23, 2016).

*The Economist* (2016), "Million-dollar babies", *The Economist*, April 2, available at: www.economist.com/news/business/21695908-silicon-valley-fights-talent-universities-struggle-hold-their (accessed November 2, 2016).

van den Bosch, A., Bogers, T. and de Kunder, M. (2016), "Estimating search engine index size variability: a 9-year longitudinal study", *Scientometrics*, Vol. 107 No. 2, pp. 839-856, doi: 10.1007/s11192-016-1863-z.

Van Noorden, R. (2014), "Online collaboration: scientists and the social network", *Nature*, Vol. 512 No. 7513, pp. 126-129, available at: www.nature.com/news/online-collaboration-scientists-and-the-social-network-1.15711

Verhaar, P. (2009), "SURE: statistics on the usage of repositories", SURF Foundation, available at: http://docplayer.net/750695-Sure-statistics-on-the-usage-of-repositories.html (accessed November 3, 2016).

Wang, Ya and Howard, P. (2012), "Google Scholar usage: an academic library's experience", *Journal of Web Librarianship*, Vol. 6 No. 2, pp. 94-108, doi: 10.1080/19322909.2012.672067.

Zeifman, I. (2015), "2015 bot traffic report: humans take back the web, bad bots not giving any ground", Incapsula Blog, December 9, available at: www.incapsula.com/blog/bot-traffic-report-2015.html (accessed June 30, 2016).

Zineddine, M. (2016), "Search engines crawling process optimization: a webserver approach", *Internet Research*, Vol. 26 No. 1, pp. 311-331, doi: 10.1108/IntR-02-2014-0045.

**Corresponding author**
Kenning Arlitsch can be contacted at: kenning.arlitsch@montana.edu