Undercounting File Downloads from Institutional Repositories

Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda

OBrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J., and Borda S. (2016) Undercounting File Downloads from Institutional Repositories. *Journal of Library Administration,*. 56(7).

Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu

COLUMN TITLE: POSIT

Column Editor: Kenning Arlitsch, Dean of the Library, Montana State University, Bozeman, MT kenning.arlitsch@montana.edu

This JLA column posits that academic libraries and their services are dominated by information technologies, and that the success of librarians and professional staff is contingent on their ability to thrive in this technology-rich environment. The column will appear in odd-numbered issues of the journal, and will delve into all aspects of library-related information technologies and knowledge management used to connect users to information resources, including data preparation, discovery, delivery and preservation. Prospective authors are invited to submit articles for this column to the editor at kenning.arlitsch@montana.edu

UNDERCOUNTING FILE DOWNLOADS FROM INSTITUTIONAL REPOSITORIES

Patrick OBrien
Semantic Web Research Director, Montana State University, Bozeman, MT USA

KENNING ARLITSCH
Dean of the Library, Montana State University, Bozeman, MT USA

Leila Sterman
Scholarly Communication Librarian, Montana State University, Bozeman, MT USA

Jeff Mixter
Software Engineer, OCLC Research, Dublin, OH USA

Jonathan Wheeler
Data Curation Librarian, University of New Mexico, NM USA

Susan Borda
Digital Technologies Development Librarian, Montana State University, Bozeman, MT USA

Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu

ACKNOWLEDGEMENT

The authors wish to express their gratitude to the Institute of Museum and Library Services, which funded this research (Arlitsch et al., 2014).

ABSTRACT

A primary impact metric for institutional repositories (IR) is the number of file downloads, which are commonly measured through third-party web analytics software. Google Analytics, a free service used by most academic libraries, relies on HTML page tagging to log visitor activity on Google's servers. However, web aggregators such as Google Scholar link directly to high value content (usually PDF files), bypassing the HTML page and failing to register these direct access events. This paper presents evidence of a study of four institutions demonstrating that the majority of IR activity is not counted by page tagging web analytics software, and proposes a practical solution for significantly improving the reporting relevancy and accuracy of IR performance metrics using Google Analytics.

KEYWORDS

Institutional repositories; IR; digital library assessment; Web analytics; Google Analytics; Log file analytics

INTRODUCTION

Institutional repositories (IR) have been under development for over fifteen years and have collectively become a significant source of scholarly content. More than 95% of the approximately 3,100 open access repositories listed in OpenDOAR are affiliated with academic institutions or research disciplines (University of Nottingham, 2016) and these repositories can add value to the research process and the reputations of institutions and their faculty. The value proposition that justifies the expense of building and maintaining open access IR is based largely on unrestricted access to their content, and on the ability of IR managers and library administrators to report impact to researchers and university administrators. Ultimately, citations may be the most valued measure of reuse and worth, and it is reasonable to expect publications to be downloaded and read before being cited. Using file download counts as a metric for scholarly value is therefore crucial for IR assessment, but it is a surprisingly difficult metric to measure accurately due to the deficiencies of web analytics tools and due to overwhelming non-human (robot) traffic.

© Patrick OBrien, Kenning Arlitsch, Leila Sterman, Jeff Mixter, Jonathan Wheeler, and Susan Borda
Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu

The scholarly information-gathering process includes a filtering approach, (Acharya, 2015) through which the researcher eventually arrives at citable scholarly content. Measurable human interaction with IR can be said to include page views or downloads of three categories:

1. **Ancillary Pages** - IR HTML pages that provide general information or navigation paths through the IR. Examples include search results, and browse pages organized by author, title, community pages, statistics, etc.
2. **Item Summary Pages -** IR HTML pages that typically include an abstract and metadata for a single scholarly work, which can help the user decide to download the full publication.
3. **Citable Content Downloads -** scholarly content that may be formally cited in the research process. These include publications, presentations, data sets, etc., accessed in a non-HTML format (i.e., .pdf, .doc, .ppt, etc.)

Current assessment practices have deficiencies that result in serious undercounting of total IR activity, leaving IR managers and stakeholders unable to accurately report on file downloads. This study examined data from four repositories: three running the DSpace platform and one running CONTENTdm. Evidence gathered from these four IR shows as much as 58% of all human-generated IR activity goes unreported by Google Analytics, the web analytics service used most frequently in academic libraries to measure use. The Research Methods and Findings sections demonstrate a pragmatic framework for reporting meaningful IR performance metrics. The data set that supports this study is available from Montana State University ScholarWorks, http://doi.org/10.15788/M2Z59N.

Research Statement

While it is possible to accurately report the first two metrics categories (Ancillary Page Views and Item Summary Page Views), Citable Content Download metrics are very difficult to report accurately. Most libraries lack the technical sophistication and resources, within their chosen web analytics methods, to identify and exclude all robot activity and to capture and report downloads generated from direct links.

Evidence presented in this study will support the following statements:

- Ancillary Page Views comprise a large portion of total IR activity being reported.
- Citable Content Downloads goes unreported by Google Analytics.
- Non-human robot activity overwhelms human activity and is too difficult to consistently filter from web analytics reports.

Literature Review

While IR content was initially defined as scholarly (Crow, 2002), some collection development policies now define the scope of IR more broadly and include institutional records and other digitized materials. This study focuses on scholarly content within IR.

Assessment of Institutional Repositories

IR assessment is acknowledged as a necessity in numerous articles in the professional literature, and is sometimes even tied to their ultimate survival. "Without understanding the significance of this service, the value of such programs may be underestimated and, consequently, funds to ensure IR survival and growth may dwindle" (Burns, Lana, & Budd, 2013). Researchers acknowledge that specific forms of measure must vary based on local needs and audience, and some assessors of IR success place less emphasis on hard metrics, noting instead that IR managers may measure their success in the comprehensiveness and growth of their repositories, and giving credence to downloads only insofar as their general ability to show "use" (Cullen & Chawner, 2010).

Most of the literature about IR assessment does focus on collecting and reporting quantitative metrics to help make the case for IR value, "Metrics for repositories can be used to provide a better understanding of how repositories are being used, which can help to inform policy decisions on future investment" (B. Kelly et al., 2012). A 2011 study of several high-profile IR reported that "assessment measures are still being developed," but that "most institutions found it easier to develop quantitative measures of success [including] the number of requests" (Campbell-Meier, 2011). Others also reinforce that specific measures based on quantifiable data will resonate, even if those reports must be customized to the audience. "By providing useful and appropriate statistics to authors, departments, the university, and other stakeholders, the library demonstrates its value as a vital partner in research, scholarship, and scholarly communication" (Bruns & Inefuku, 2016). Bruns and Inefuku count item downloads among a number of metrics that should be assembled based on institutional mission and on audience. Lagzian et al. list the ability of the system to make "available the number of downloads and views of full text files" as one of the top critical success factors for IR (Lagzian, Abrizah, & Wee, 2015).

Despite the recognition that quantifiable metrics, including downloads, are useful, there is evidence that data and reporting abilities for IR are lacking. "While libraries determine the most appropriate benchmark for success for their respective IRs, the need for more precise usage data will be central to assessment efforts" (Fralinger & Bull, 2013). The message conveying the importance of assessment is also not necessarily widely accepted. In a 2013 publication that surveyed the usage of U.S.-based IR by international audiences, Fralinger and Bull report that many IR administrators "seem to be unaware, apathetic, or unprepared to do IR assessment, specifically from an international perspective" (Fralinger & Bull, 2013).

Some prior research has already pointed out that discrepancies exist in download reports when search engines send users directly to the file. A 2006 study at the University of Wollongong, Australia, noted that its Digital Commons repository statistics suggested that users accessing [the IR] from Google are in the majority of cases going straight to the document pdf, rather than to the cover page" (Organ, 2006).

Overview of Web Analytics Methods

Reporting visitation and use of websites and digital repositories is made possible through the use of web analytics software, which may be divided into two classes: 1) page tagging; and 2) log file. Brief descriptions of these types follow, but more in-depth analyses are available in other studies (Clifton, 2012; Fagan, 2014; Jansen, 2006; Nakatani & Chuang, 2011).

Page tagging analytics

This class of analytics software is typically delivered as Software as a Service (SaaS), where the software package usually resides on the vendor's servers. Popular page tagging software include free packages such as Google Analytics, and costly options such as WebTrends and Adobe Marketing Cloud. Page tagging analytics relies on a piece of tracking code (usually JavaScript) that is embedded on each HTML page of the website in question. The tracking code is keyed to the account holder and acts as a beacon to the software package on the vendor's servers. A display of the HTML page triggers a signal from the tracking code to the software package, where the visit is registered along with various other pieces of information that can include the referral site, search terms, user's geographical location, type of device, etc.

Log file analytics

Log file analytics software provide reports on the data normally collected by server logs. This type of software is typically installed and managed locally by server administrators, and "web log analysis software ... then can be used to analyze the log file" (Nakatani & Chuang, 2011). Log file analytics software includes the packages built into DSpace and ePrints, as well as other packages such as WebLog Expert.

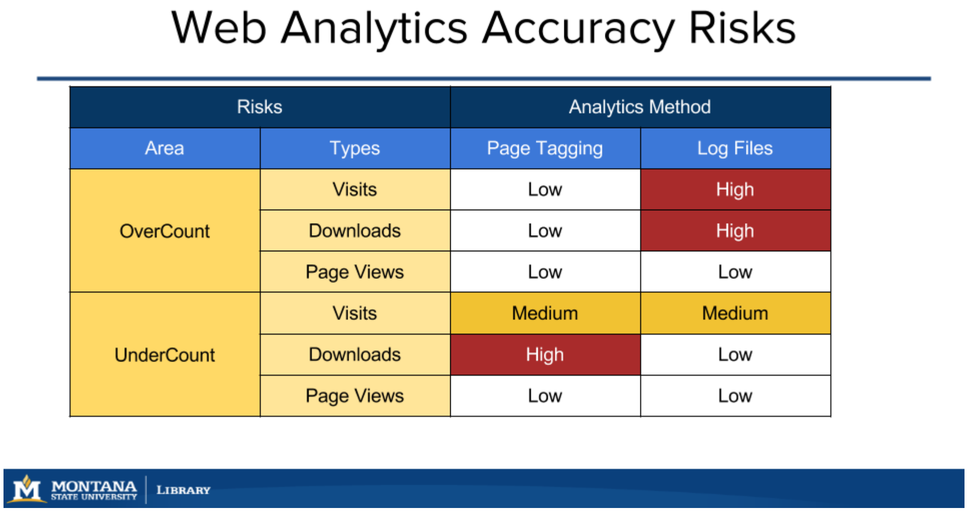
Which Type is Better?

Both classes of analytics software have strengths and weaknesses. Correctly configured, page tagging analytics software can provide a holistic view of all the organization's web properties, including the ability to see the paths that users follow through a domain. Sophisticated reports can be generated using tools built into the software, and in a SAAS environment there is no need for local updates or maintenance of the software itself.

Log file analytics can provide very granular information about IR activity, but since the software is managed locally it can impose a small administrative burden. Log file analysis can be difficult to configure if aggregating data from more than one physical webserver; manual compilation of reports is required when multiple servers comprise the website of an organization. On the other hand, a distinct advantage of log file analysis is that user and institutional information is not shared with a third party. Over the past few years, analytics plug-ins have been developed for popular file index stacks, such as Solr and Elasticsearch.

Both page tagging and log file analytics carry significant risks for inaccurate reporting of IR activity (see Table 1). Page tagging analytics software carries high risk for undercounting non-HTML file downloads, particularly when users are referred directly to the file from an external source (see Figure 1). Log file analytics software, on the other hand, carries a high risk of over-counting due to the dynamic "cat and mouse" game required to identify and filter bots, crawlers and scrapers. In websites that see fewer than 10,000 visitors per day it is estimated that less than 30% of online traffic is human-initiated (Zeifman, 2015). Paradoxically, log files can also sometimes underestimate activity due to proxy and browser caching (Ferrini & Mohr, 2009).

Although web analytics can help report IR activity, there is a significant amount of academic paper sharing that may never be tracked. It is nearly impossible, within the varied scholarly communication ecosystem, to capture all the interactions that exist with any given paper.

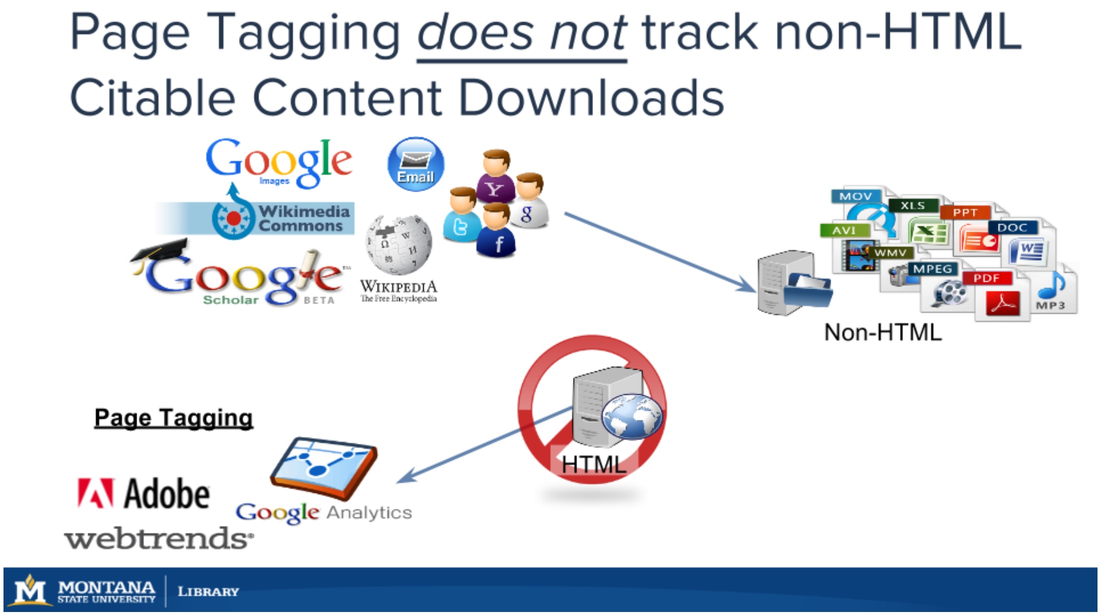


Web Analytics Accuracy Risks

Risks		Analytics Method	
Area	Types	Page Tagging	Log Files
OverCount	Visits	Low	High
	Downloads	Low	High
	Page Views	Low	Low
UnderCount	Visits	Medium	Medium
	Downloads	High	Low
	Page Views	Low	Low

MONTANA STATE UNIVERSITY
LIBRARY

TABLE 1: RISKS ASSOCIATED WITH EACH TYPE OF WEB ANALYTICS METHOD



Page Tagging *does not* track non-HTML Citable Content Downloads

FIGURE 1: PAGE TAGGING ANALYTICS DOES NOT TRACK CITABLE CONTENT DOWNLOADS

The Problem of Robots

Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu

Undercounting is the basis for this research, but over-counting non-human activity is also a concern. Log file analytics store every request for every page and file. Robots (bots) create a large bias in log-based analytics because they account for almost 50% of all internet traffic (Zeifman, 2015) and over 85% of IR downloads (Information Power Ltd, 2013). While DSpace has a bot filtering feature, it only addresses known bots, which fall under the "good bots" category and include crawlers from Google or Bing whose job it is to index IR content. "Bad bots" on the other hand are used for malicious purposes, such as probing for server vulnerabilities that can be used to infect visitors, generate SEO referral spam, or harvest the entire IR content to generate traffic on other websites. While "good bots" are easily detected and screened from reports, they account for only 40% of total bot activity. Log-based analytics methods have difficulty in effectively identifying and excluding "bad bot" activity that accounts for the other 60% of total bot activity (Zeifman, 2015). The problem of bots skewing reports has led to development of "more sophisticated – but practical – algorithms to improve filtering that will eventually become incorporated into the COUNTER standard" and will be used to help measure use and impact of IR (MacIntyre & Jones, 2016). However, until these sophisticated solutions are available, using Google Search Console Clicks and Google Analytics Events, as described in the Research Methods section, may be the most accurate for reporting IR downloads.

Google Analytics

Although some researchers argue that Google Analytics is inappropriate for educational use, since it was built for e-commerce rather than an educational environment (Dragos, 2011), our related research has shown that the majority of academic libraries still use this tool. In a study on privacy that will be published in 2017, we found the presence of Google Analytics tracking code in over 80% of the 263 academic libraries we surveyed. Outside the realm of academic libraries, it has been reported that "more than 60% of all websites on the Internet use Google Analytics, Google AdSense or another Google product using tracking beacons" (Hornbaker & Merity, 2013; Piwik development team, 2016).

Google Analytics provides a very accurate metric for determining the number of HTML pages viewed by humans. Most bots are incapable of running the Google Analytics JavaScript tracking code needed to register a page view. Google's primary business model is digital advertising and companies use their analytics software to maximize eCommerce profit. As a result, Google has a vested interest in ensuring that only human activity is tracked and reported. Given this emphasis, it is not worth the effort for libraries to spend money or staff time to meticulously eliminate bot activity through their own local system. The tools and infrastructure provided by Google Analytics and the Google Search Console API are the most cost-effective, although they come with legitimate privacy concerns.

Standard configuration of Google Analytics provides statistics on HTML page views, but additional configuration called “event tracking” (Bragg et al., 2015) is required to track non-HTML Citable Content Downloads that comprise the bulk of citable IR content. Other researchers have previously noted the difficulty of tracking non-HTML downloads in Google Analytics: “Without implementing event tracking, Google Analytics has no way to track these [PDF] downloads, and the data will not be included in any reports” (Farney & McHale, 2013) and “direct downloads of PDFs hosted in repositories may not be reported unless Google Analytics has been configured appropriately,” resulting in underestimates (Brian Kelly, 2012). In a related study, Kelly and co-authors describe applying GA tracking code to the download link on the HTML page (B. Kelly et al., 2012), but this method doesn’t address visitors who arrive directly to the PDF. Burns, Lana, and Budd also write that “web log analytics may under report IR use” and refer to a DSpace solution developed at the University of Edinburgh, which “created a redirect so a user's click on a PDF link in a Google search results list will take the user to the file's item page in the IR, rather than directly to the file” (Burns et al., 2013). Claire Knowles’ slide presentation shows large increases in download statistics once the redirect was put into place in December 2011 (Knowles, 2012). However, this solution does not seem to have gained widespread traction, and it is unlikely that Google and Google Scholar would look favorably upon a redirect when those search engines offer direct links to the files. Similarly, DSpace's current 5.x implementation has a feature to track download events within Google Analytics. However, after reviewing the DSpace code we determined the current method relies on the Google Analytics API – not Google Search Console API- and, thus, is limited to tracking file downloads that originate from a DSpace HTML page. We also confirmed this by isolating high-use non-HTML files and comparing Google Analytics Download Events with Google Search Console Clicks.

Finally, it should be noted that vocabulary plays a role in measuring and communicating impact through web analytics. The library science profession has long referred to digital objects (such as PDF files in an IR) as "items" (Lagoze, Payette, Shin, & Wilper, 2006; Tansley et al., 2003), while Google Analytics calls all HTML pages, including those that contain abstracts and metadata "items." This can cause confusion when communicating impact. As we noted at the start, we refer to pages containing only metadata and abstracts as Item Summary Pages, and while technically they contain all the information required for citation, one hopes that a scholarly citation would only result from the download and reading of the full publication, (i.e., what we call the Citable Content Downloads).

RESEARCH METHODS

The data set for this study was collected from four institutional repositories, whose activity was monitored during a 134-day period during spring academic semester in 2016:

- LoboVault - University of New Mexico - https://repository.unm.edu
- MacSphere - McMaster University - https://macsphere.mcmaster.ca
- ScholarWorks - Montana State University – http://scholarworks.montana.edu
- USpace - University of Utah – http://uspace.utah.edu

The first three repositories run the DSpace platform, while USpace at the University of Utah runs CONTENTdm. The University of New Mexico is in the process of migrating to a Digital Commons platform, and expects to go live before end of summer, 2016. Data for this study were collected from UNM's DSpace platform.

TOOLS

Data were gathered and compared using a number of tools and configurations. Google Analytics, deployed in conjunction with the Google Search Console (previously known as Webmaster Tools), was used to help compile activity, and DSpace usage statistics and Solr stats were also utilized. The following list explains which specific activity was pulled from each tool:

1. Google Analytics
   a. Page Views
   b. Events
2. Google Search Console API
   a. Clicks
3. DSpace
   a. Google Analytics Statistics
   b. Usage Statistics
   c. Solr Stats
   d. Solr Item Metadata

GOOGLE ANALYTICS

In order to exclude activity that does not support the mission of open access IR and to identify search referral details, we applied IP address filters that excluded library staff activity, added the Organic Search Source setting to identify referral detail about Google, Google Scholar and Google Image, and enabled bot filtering.

The Google Analytics reports used for this study are listed, below. PV=Page Views.

1. Total IR HTML PV - estimated using Google Analytics > Behavior > Site Content > All Pages Report. Reported PV were used in lieu of Unique PV to ensure our study findings were conservative (Google 2016b).
2. Total IR Item Summary PV - estimated by refining the report listed in #1 with an Advanced Filter using Regular Expressions. Regular Expressions were developed to exclude any activity involving HTML Ancillary Pages for the unique configuration in each IR.
3. Ancillary PV - estimated by Total IR HTML PV less Total Item Summary PV
4. Download Events - estimated using Google Analytics > Events > Behavior > Events > Pages report. Note: only Montana State University and the University of Utah IR had configured their IR software and Google Analytics to track Events, as will be seen in Table 2.

Google Search Console API

Google Search Console provides the count of human clicks that each URL receives from the search results pages (SERP) of Google's search properties (Google, Inc., 2016). However, account holders can only access the last 90 days of visitation data via the Google Search Console interface. To accumulate persistent data for our study, we used Python scripts to access the Google Search Console API to extract URLs that received one or more clicks each day. We then applied the Regular Expressions developed for estimating Total IR Item Summary PV, above. We also included rules to include only URLs for non-HTML files in this estimate. This method allowed us to retrieve every record on every day from Google Search Console, with no limitations. In brief, we were able to extract a persistent dataset with the granular detail required for this study.

DSpace

A previous researcher had asserted that the reported DSpace statistics may be biased by as much as 85% over-counting due to bot activity (Greene, 2016). We corroborated this claim by temporarily bypassing the local host restriction (Masár, 2015) and acquiring the detailed download records from the Solr statistics core. These records were then joined with the metadata records from the Solr search core (Diggory & Luyten, 2015) in the Montana State and University of New Mexico IR. We also tried to analyze the DSpace Google Analytics Statistics feature, but learned of a current bug (Dietz, 2015) in DSpace 5.x that prevented DSpace from generating those statistics for our study participants' IR Items.

Data Set

The resulting dataset (OBrien et al., 2016) contains over 57,087 unique URLs in 413,786 records that received one or more human clicks via Google SERP from January 5, 2016 to May 17, 2016 (134 days). Using the Google Search Console API, we were able to determine the total number of invisible Citable Item Downloads (item downloads that did not originate from the IR website and were not reported in Google Analytics – see Figure 4). After aggregating these data, a regular expression was used to exclude URLs containing non-scholarly material, such as collection landing pages. This resulted in a set of data that could be used to determine the number of non-HTML files (.pdf, .jpeg, MS Word documents, MS PowerPoint, .txt datasets, MS Excel files, etc.) that were directly downloaded from Google SERP.

LIMITATIONS

This study involves only four repositories, although the compiled data set includes over 400,000 URLs. The data were gathered during the height of the spring semester when classes were in session, a time during which use of the IR at the four universities should have been high. However, it could be argued that the fall semester might have garnered more activity, and ideally, an entire year of data would be collected and analyzed for a larger number of IR. As with any study that gathers data from a dynamic environment, the data should be considered a snapshot in time.

Another limitation is that only two repository software platforms (DSpace and CONTENTdm) are represented in this study. DSpace is by far the most widely used IR software in the world and its selection is justifiable on that basis. While CONTENTdm has seen broad adoption in cultural heritage digital libraries, it is not very widely used as an IR platform. However, gathering data from IR is contingent on relationships that provide a specific level of access, and the CONTENTdm repository was another data set to which the authors had access. A larger study should ideally include other platforms, such as Digital Commons and ePrints.

Finally, Google Search Console brings value to this study by helping to include tracking of non-HTML downloads. However, its ability to count downloads is limited to clicks that originate from other Google properties, and therefore some number of direct downloads from non-Google properties have been missed in this study.

FINDINGS

The total IR activity from the four repositories that we can report with a high level of confidence and accuracy, was calculated by combining Google Analytics Page Views, Google Search Console Clicks, and Google Analytics Events. Evidence gathered from the IR in this study show as much as 58% of all human-generated IR activity goes unreported by Google Analytics, the web analytics tool used most frequently in academic libraries to measure use.

Table 2 shows the Total IR Activity that was collected from the four repositories during the 134-day data-collection period. HTML Item Summary Page Views and Ancillary Page Views combined to provide the total number that Google Analytics was able to report for each of the four repositories (see Total Google Analytics HTML PV). Only Montana State University and the University of Utah had configured their IR and Google Analytics to report Download Events, which explains why there are no data for the McMaster University and University of New Mexico repositories. The figures for Citable Content Downloads (last column) was extracted from Google Search Console API. These downloads were in addition to the Total Google Analytics HTML PV figures, demonstrating rather dramatically how much high-value activity Google Analytics is unable to capture.

IR	Item Sumary PV	Ancillary PV	Total Google Analytics HTML PV	Download Events	Citable Content Downloads
scholarworks.montana.edu	26,735	23,350	50,085	7,129	77,380
macsphere.mcmaster.ca	51,150	71,585	122,735	n/a	133,342
repository.unm.edu	83,491	59,289	142,780	n/a	166,320
content.lib.utah.edu	122,927	47,569	170,496	19,226	159,536

TABLE 2: TOTAL ACTIVITY FROM FOUR IR DURING THE 134-DAY TEST PERIOD

is a graphical representation of the Total Google Analytics HTML PV figures shown for each repository in Table 2

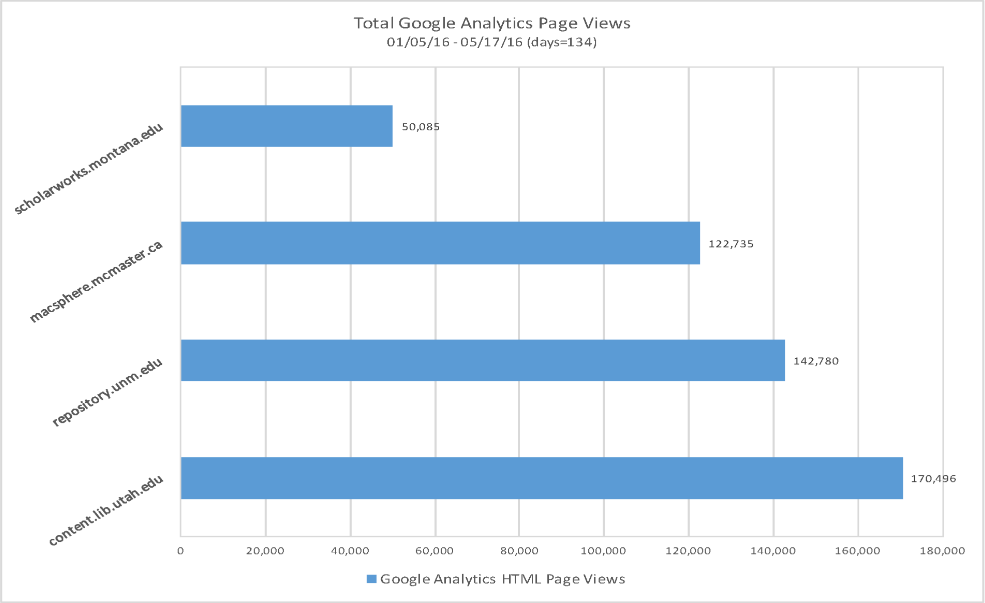


FIGURE 2: CHART REPRESENTATION OF TOTAL GOOGLE ANALYTICS HTML PAGE VIEWS FROM THE FOUR REPOSITORIES TRACKED FOR THE STUDY

Figure 3 shows the total IR page views that were reported via Google Analytics for the four repositories, and each set of results is categorized here as Ancillary Page Views and Item Summary Page Views. The range of Ancillary Page Views across the four repositories was 28% - 58%, for a weighted average of 41.51% Ancillary PV. As explained earlier, Ancillary Pages are the low-value HTML pages that provide general information or navigation paths through the IR, while the Item Summary Pages contain abstracts and metadata for a single scholarly work.

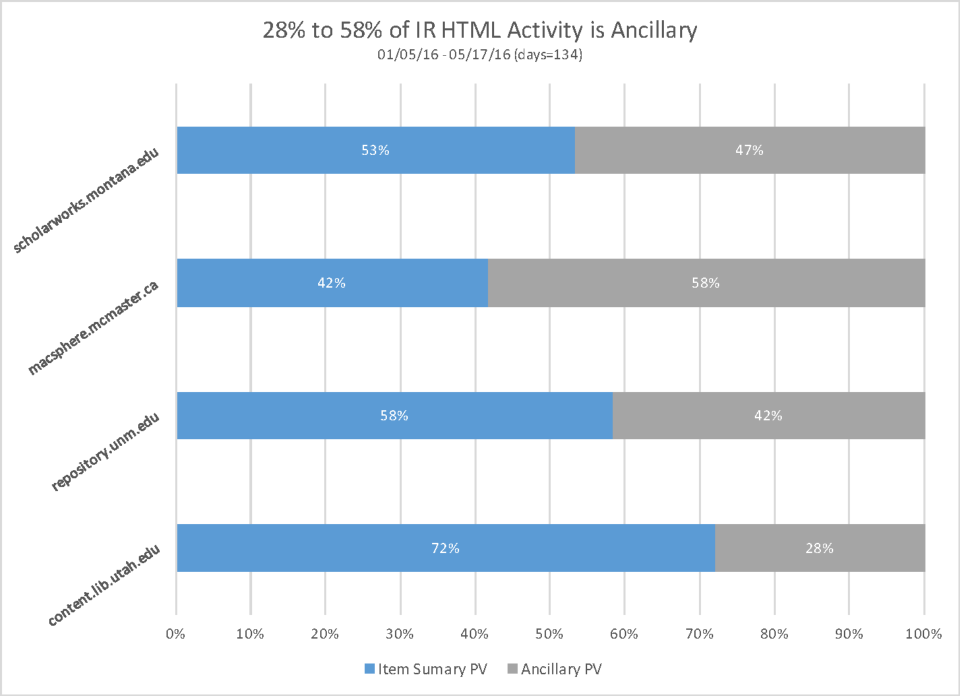


FIGURE 3: PERCENT OF ITEM SUMMARY AND ANCILLARY PAGE VIEWS (PV)

Figure 4 shows the total IR activity that our study identified via Google Analytics and Google Search Console. 46% - 58% of the activity was invisible via Google Analytics, with a weighted average of 51.1% of IR activity being invisible without the use of Google Search Console Clicks.

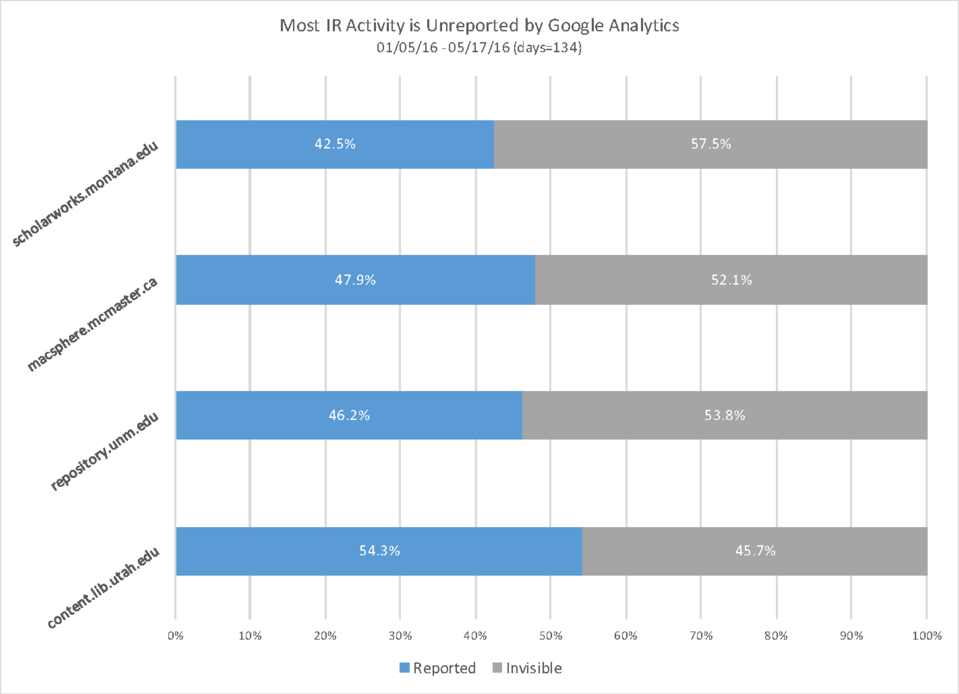


FIGURE 4: UNREPORTED IR ACTIVITY IN GOOGLE ANALYTICS

Discussion

The true value of an IR is contained within its research papers and data sets, and we refer to measureable interaction with these files as Citable Content Downloads. Our study demonstrates that the most popular analytics methods miss or underreport this most important metric of IR activity. The analytics reporting methods we introduced in this study provide a framework for more accurate measurement of IR activity.

There are several paths that a user may take to reach citable content, which is most often a PDF or other type of non-HTML file. A visit may be directed from a known link, a web search service, or by browsing the IR itself. In the case of the first two paths the user is often linked directly, bypassing the HTML pages of the repository and arriving immediately at the desired content. The third path involves direct use of the IR website, through which the user may eventually land on the HTML Item Summary Page (containing abstract and metadata), and from there s/he may click on the link on that page that downloads the non-HTML file. There is no guarantee that a user will open the file after it has been downloaded, but s/he certainly could not read the publication or work with the data set prior to download. Together, Item Summary Page Views (collected through Google Analytics), and Citable Content Downloads (collected using the Google Search Console API), are excellent indicators of IR impact that may predict eventual citation activity.

Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu

Data from this study show enormous numbers of high-value Citable Content Downloads being missed by Google Analytics. For example, the data from Table 2 show that Google Analytics failed to report 100% of Citable Content Downloads at McMaster University and the University of New Mexico, while 91.6% were missed at Montana State University and 89.2% at the University of Utah. In addition to these large reporting errors, there are still Citable Content Downloads being missed due to their origination outside the Google ecosystem. We believe that our tested methods produce a highly accurate picture of IR activity that is as granular as the IR activity reported in server logs, however, our methods also have the added convenience of being pre-filtered for bots and other non-human activity. More research and analysis is required to determine exactly how much human activity goes unaccounted, but our preliminary estimates indicate the activity we cannot accurately measure is small and may have little effect on reporting.

Event tracking is an added configuration that should ideally be implemented in Google Analytics, but its effect may be overrated. Our analysis showed that event tracking accounted for only 8%-11% of total Citable Content downloads in our study.

Adding Google Search Console results improves accuracy by providing the number of non-HTML downloads, but those represent only clicks that originated from a Google search property. Therefore, this study still may be missing a significant number of Citable Content Downloads originating from other sources that bypass HTML pages in the IR. These other sources may include Bing, Yahoo!, Wikipedia, numerous social media sites such as FaceBook, Twitter, Reddit, CiteULike, professional and academic sites like LinkedIn, ResearchGate, Academia.edu, Mendeley, and direct email referrals. Publishers have similar data capture limitations beyond their journal web pages, and they track use through services like DataCite that produce and monitor Digital Object Identifiers (DOI) (Paskin, 2000). Publishers may go one step further by working through controlled services like ReadCube (Goncharoff, 2014), but that practice limits access to devices that are set up for the ReadCube application and runs counter to the IR mission of facilitating access to scholarship rather than limiting it.

The source of a web referral to an IR matters. From an institutional perspective, a visitor referred by Google Scholar carries greater value than one referred from Google, Facebook or Yahoo!. Google Scholar users primarily represent researchers seeking scholarly publications, and they are more likely to download IR files and use them to support their own research. This is a high-value audience that should be of great interest to IR managers. Metadata and site crawling problems have historically limited many IR from establishing the kind of robust relationship with Google Scholar that facilitates consistent and accurate harvesting of publications (Arlitsch & O'Brien, 2012). But data collected in the current study indicate that repositories indexed by Google Scholar receive 48%-66% of their referrals from Google Scholar. These figures are significant and imply that a good relationship with Google Scholar is worthwhile, as it leads to considerable high-quality interactions.

CONCLUSION

Most IR managers are only able to measure a small part of the high-value traffic to their IR; much of what they see is indicative of visits to the site's low-value HTML pages rather than a measure of citable content downloads. The standard configuration of the most widely-used analytics service in academic libraries, Google Analytics, fails to capture the vast majority of non-HTML Citable Content Downloads from IR. This seriously limits the effectiveness of IR managers and library administrators when they try to make the case for IR usefulness and impact.

Current mechanisms for collecting accurate analytics are limited and further study is warranted, but the methods tested in this study are promising. Some repository platform developers make claims that they address analytics data collection and reporting, but it can be difficult to know exactly which techniques are being applied and how effective they are, particularly in a proprietary environment. For example, simply using a bot list isn't good enough because "bad bots" are constantly changing to avoid detection and what worked yesterday may not work tomorrow.

We have the potential to know so much more about the movement and use of research than we did when bound paper copies circulated from office to office, but our knowledge will only increase if the tools are set up appropriately and calibrated for the task. Trying to capture currently invisible activity is a large endeavor, but one that will help us make a stronger use case for IR, better understand IR user needs, and continue to improve access to research. The ability to report downloads can be a powerful tool to help faculty engage with IR. Citations may take years to appear in the literature, but repository downloads act as a proxy measure, giving the IR manager a more immediate and understanding of use.

REFERENCES

Acharya, A. (2015, September). *What happens when your library is worldwide and all articles are easy to find?* Videorecording presented at the Association of Learned and

Professional Society Publishers, London. Retrieved from https://www.youtube.com/watch?v=S-f9MjQjLsk

Arlitsch, K., OBrien, P., Kyrillidou, M., Clark, J. A., Young, S. W. H., Mixter, J., … Stewart, C. (2014). *Measuring Up: Assessing Accuracy of Reported Use and Impact of Digital Repositories* (Funded grant proposal) (pp. 1–10). Institute of Museum and Library Services. Retrieved from http://scholarworks.montana.edu/xmlui/handle/1/8924

Arlitsch, K., & O'Brien, P. S. (2012). Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech*, *30*(1), 60–81. http://doi.org/10.1108/07378831211213210

Bragg, M., Chapmen, J., DeRidder, J., Johnston, R., Junus, R., Kyrillidou, M., & Stedfeld, E. (2015). *Best practices for Google Analytics in digital libraries*. Digital Library Federation. Retrieved from https://docs.google.com/document/d/1QmiLJEZXGAY-s7BG_nyF6EUAqcyH0mhQ7j2VPpLpxCQ/edit

Bruns, T., & Inefuku, H. W. (2016). Purposeful metrics: matching institutional repository metrics to purpose and audience. In *Making Institutional Repositories Work* (pp. 213–234). West Lafayette, IN: Purdue University Press. Retrieved from http://lib.dr.iastate.edu/digirep_pubs/4/

Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional Repositories: Exploration of Costs and Value. *D-Lib Magazine*, *19*(1/2). http://doi.org/10.1045/january2013-burns

Campbell-Meier, J. (2011). A Framework for Institutional Repository Development. In D. E. Williams & J. Golden (Eds.), *Advances in Library Administration and Organization* (Vol. 30, pp. 151–185). Emerald Group Publishing Limited. Retrieved from http://www.emeraldinsight.com/doi/abs/10.1108/S0732-0671%282011%290000030006

Clifton, B. (2012). *Advanced Web metrics with Google Analytics* (3rd ed). Indianapolis, Ind: Wiley.

Crow, R. (2002). *The case for institutional repositories: a SPARC position paper*. (ARL Bimonthly Report No. 223). Retrieved from http://works.bepress.com/ir_research/7

Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu

Cullen, R., & Chawner, B. (2010). Institutional repositories: assessing their value to the academic community. *Performance Measurement and Metrics*, *11*(2), 131–147. http://doi.org/10.1108/14678041011064052

Dietz, P. (2015, November 8). Google Analytics Statistics not relating parent comm/coll to bitstream download. *DuraSpace*. Retrieved from https://jira.duraspace.org/browse/DS-2899

Diggory, M., & Luyten, B. (2015, August 21). SOLR Statistics - DSpace 5.x Documentation - DuraSpace Wiki [Wiki]. Retrieved July 1, 2016, from https://wiki.duraspace.org/display/DSDOC5x/SOLR+Statistics#SOLRStatistics-WebUserInterfaceElements

Dragos, S.-M. (2011). Why Google Analytics cannot be used for educational web content. In *Next Generation Web Services Practices (NWeSP)* (pp. 113–118). Salamanca: IEEE. http://doi.org/10.1109/NWeSP.2011.6088162

Fagan, J. C. (2014). The Suitability of Web Analytics Key Performance Indicators in the Academic Library Environment. *The Journal of Academic Librarianship*, *40*(1), 25–34. http://doi.org/10.1016/j.acalib.2013.06.005

Farney, T., & McHale, N. (2013). *Maximizing Google Analytics: six high-impact practices* (Vol. 49). Chicago, IL: ALA TechSource.

Ferrini, A., & Mohr, J. J. (2009). Uses, Limitations, and Trends in Web Analytics. In *Handbook of Research on Web Log Analysis* (pp. 122 – 140). IGI Global. Retrieved from http://www.igi-global.com/chapter/uses-limitations-trends-web-analytics/21999

Fralinger, L., & Bull, J. (2013). Measuring the international usage of US institutional repositories. *OCLC Systems & Services: International Digital Library Perspectives*, *29*(3), 134–150. http://doi.org/10.1108/OCLC-10-2012-0039

Goncharoff, N. (2014, December 10). Clearing Up Misperceptions About Nature.com Content Sharing [News Blog]. Retrieved July 7, 2016, from https://www.digital-science.com/blog/news/clearing-up-misperceptions-about-nature-com-content-sharing/

Google, Inc. (2016). Search Analytics Report. Retrieved April 1, 2016, from https://support.google.com/webmasters/answer/6155685

Greene, J. (2016). Web robot detection in scholarly Open Access institutional repositories. *Library Hi Tech*, *34*(3). Retrieved from http://hdl.handle.net/10197/7682

Hornbaker, C., & Merity, S. (2013). Measuring the impact of Google Analytics: Efficiently trackling Common Crawl using MapReduce & Amazon EC2. Retrieved from http://smerity.com/cs205_ga/

Information Power Ltd. (2013). *IRUS download data – identifying unusual usage* (IRUS Download Report). Retrieved from http://www.irus.mimas.ac.uk/news/IRUS_download_data_Final_report.pdf

Jansen, B. J. (2006). Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research*, *28*(3), 407–432. http://doi.org/10.1016/j.lisr.2006.06.005

Kelly, B. (2012, August 29). MajesticSEO analysis of Russell Group University repositories. Retrieved from http://ukwebfocus.com/2012/08/29/majesticseo-analysis-of-russell-group-university-repositories/

Kelly, B., Sheppard, N., Delasalle, J., Dewey, M., Stephens, O., Johnson, G., & Taylor, S. (2012). Open metrics for open repositories. In *OR2012: the 7th International Conference on Open Repositories*. Edinburgh, Scotland. Retrieved from http://opus.bath.ac.uk/30226/

Knowles, C. (2012). *Surfacing Google Analytics in DSpace*. Presented at the Open Repositories, Edinburgh, Scotland. Retrieved from http://or2012.ed.ac.uk/?s=Knowles&searchsubmit=

Lagoze, C., Payette, S., Shin, E., & Wilper, C. (2006). Fedora: an architecture for complex objects and their relationships. *International Journal on Digital Libraries*, *6*(2), 124–138. http://doi.org/10.1007/s00799-005-0130-3

Lagzian, F., Abrizah, A., & Wee, M. C. (2015). Critical success factors for institutional repositories implementation. *The Electronic Library*, *33*(2), 196–209. http://doi.org/10.1108/EL-04-2013-0058

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. http://doi.org/10.1080/0361526X.2016.1148423

Masár, I. (2015, December 11). Solr - DSpace - DuraSpace Wiki [Wiki]. Retrieved July 1, 2016, from https://wiki.duraspace.org/display/DSPACE/Solr#Solr-Bypassinglocalhostrestrictiontemporarily

Nakatani, K., & Chuang, T. (2011). A web analytics tool selection method: an analytical hierarchy process approach. *Internet Research*, *21*(2), 171–186. http://doi.org/10.1108/10662241111123757

OBrien, P., Arlitsch, K., Sterman, L., Mixter, J., Wheeler, J., & Borda, S. (2016). *Data set supporting the study "Undercounting file downloads from institutional repositories."* en, Bozeman, MT: Montana State University. Retrieved from http://scholarworks.montana.edu/xmlui/handle/1/9939

Organ, M. (2006). Download Statistics - What Do They Tell Us?: The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia. *D-Lib Magazine*, *12*(11). http://doi.org/10.1045/november2006-organ

Paskin, N. (2000). E-citations: actionable identifiers and scholarly referencing. *Learned Publishing*, *13*(3), 159–166. http://doi.org/10.1087/09531510050145308

Piwik development team. (2016). Web Analytics Privacy in Piwik. Retrieved June 13, 2016, from http://piwik.org/privacy/

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., & Smith, M. (2003). The DSpace institutional digital repository system: current functionality. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 87–97). Los Alamitos, Calif.: IEEE Computer Society. Retrieved from http://dspace.mit.edu/bitstream/handle/1721.1/26705/Tansley_2003_The.pdf?sequence=1

University of Nottingham. (2016, July 14). Open access repository types - worldwide. Retrieved from http://www.opendoar.org/onechart.php?cID=&ctID=&rtID=&clID=&lID=&potID=&rSoftWareName=&search=&groupby=rt.rtHeading&orderby=Tally%20DESC&charttype=pie&width=600&height=300&caption=Open%20Access%20Repository%20Types%20-%20Worldwide

Zeifman, I. (2015, December 9). 2015 Bot Traffic Report: Humans Take Back the Web, Bad Bots Not Giving Any Ground. Retrieved from https://www.incapsula.com/blog/bot-traffic-report-2015.html

Address correspondence to Patrick OBrien, Semantic Web Research Director, Montana State University,
P.O. Box 173320, Bozeman, MT 59717-3320, USA. E-mail: patrick.obrien4@montana.edu