• OPEN REPOSITORIES REPORT SERIES

Institutional Repository Statistics: Reliable, Consistent Approaches for Canada

by *Will Roy, Brian Cameron, and Tim Ribaric*
(members of the CARL Open Repositories Working Group's Task Group for Standards for IR Usage Data)

DECEMBER 2019

www.carl-abrc.ca

CARL ABRC
CANADIAN ASSOCIATION OF RESEARCH LIBRARIES • ASSOCIATION DES BIBLIOTHÈQUES DE RECHERCHE DU CANADA

Table of Contents

Introduction

> "Usage metrics are an effective way for libraries to demonstrate the value of their institutional repositories, however, existing tools are not always reliable and can either undercount or overcount file downloads. As well, although statistics can sometimes be accessed through the various repository interfaces, without an agreed standard it is impossible to reliably assess and compare usage data across different IRs in any meaningful way."[1]

The Task Group for Standards for IR Usage Data has undertaken an information-gathering exercise to better understand both the existing practices of Canadian repositories, as well as the emerging tools and processes available for repositories to track and monitor usage more effectively. This exercise directly links to the broader goals of the Open Repositories Working Group, which are to "strengthen and add value to the network of Canadian open access repositories by collaborating more closely and adopting a broader range of services."[2]

Our recommended course of action is for all Canadian IRs to collectively adopt OpenAIRE Statistics. This path aligns with the following recommendations which our group also advances:

Recommendations

We suggest the following **Mandatory (M)** and **Optional (O)** recommendations:

> *R1(M): All Canadian IRs should adopt the COUNTER Code of Practice.*
>
> *R2(M): All Canadian IRs should select a service that allows for interoperability with other web services via a fully open, or accessible, permissions-based API.*
>
> *R3(M): All Canadian IRs should use a statistics service that practices transparent communication and maintains a governance strategy.*

In addition, we strongly urge for the future that Canadian IRs consider the following advice.

> *R4(O): Make further investments into understanding and utilizing the common log format (CLF).*

[1] Task Group for Standards for IR Usage Data (CARL Open Repositories Working Group) terms of reference, 2018. http://www.carl-abrc.ca/wp-content/uploads/2019/06/ORWG-task-group-ToR-june2019.pdf.

[2] CARL Open Repositories Working Group terms of reference, 2018. http://www.carl-abrc.ca/wp-content/uploads/2018/03/ToR_ORWG_eng.pdf.

R5(O): Conduct research into the privacy implications of collecting use statistics via third party services with commercial interests and consider available alternatives.

R6(O): Practice a healthy skepticism towards tools and solutions that promise "increased" usage statistics, and instead advocate for responsible collection assessment based on multiple aspects of use.

The following table shows how three existing services have been scored based on the desired features included in the recommendations above.

RECOMMENDATION	IRUS-UK	OpenAIRE Statistics	RAMP
#1 Counter compliance	3	3	1
#2 Interoperability (i.e. API access, open data, dashboards, etc.)	3	3	2
#3 Governance / transparency	3	3	3
#4 Can utilize data from common log format	N/A	N/A	N/A
#5 Privacy focus	3	3	2
#6 Focus on "accuracy" of results vs. "quantity" of results	3	3	2
Special category: Bot handling	3	3	2
Special category: Ease of setup for Canadian IRs	2	2	3
Total Score	**20**	**20**	**15**

Legend

3 = Favourable approach

2 = Acceptable approach

1 = Needs work

N/A = Not Applicable to scoring because recommendation is for repository managers and not a feature

Further context for these recommendations is provided in the subsequent report of findings.

Report of Findings

> *"By collecting and presenting various metrics of repository usage, repository managers are able to offer a valued service to both researchers and institutions."*[3]

Background Context

This report explores usage statistics as a metric worthy of interest and capture for repository managers and is aimed at identifying the means by which to achieve that goal. Repositories operate within a diverse scholarly ecosystem with many players; therefore, it is worthwhile to consider the context in which metrics can be of value to the broader community of stakeholders.

Defining and Differentiating Usage Statistics

Usage statistics, in this context, refer to the capture of views and downloads of a specific item hosted on either a publishing platform or other digital distribution platform such as a repository. There are two common methods for capturing usage statistics: log file analysis, which tracks usage events on the server-side, and page tagging, which tracks usage events on the client side.[4] Usage statistics reflect usage originating from one particular access point, though there have been ongoing attempts to display and combine usage statistics happening at multiple access points as seen through projects such as the Distributed Usage Logging project[5] by COUNTER and CrossRef, and the Public Library of Science's Project Lagotto tool.[6]

Usage statistics reflect a different coverage of users than citation-based indicators. Unlike citation metrics, which are based on the activities of authors, "there are many potential users (students, policy makers, the interested public) who read publications or use data without ever publishing. In addition, not everything a researcher reads is referenced in her publications." [7] The limitations of citation-based metrics were demonstrated by the Public Library of Science in which they reported that, "only about one in 70 users who download a PDF of the paper will cite it".[8] Considering the

[3] Confederation of Open Access Repositories (COAR). 2013. Incentives, integration, and mediation: Sustainable practices for populating repositories. Retrieved from: https://www.coar-repositories.org/files/Sustainable-best-practices_final2.pdf.

[4] COUNTER. 2019. The COUNTER code of practice for release 5. Retrieved from: https://www.projectcounter.org/code-of-practice-five-sections/6-logging-usage/.

[5] https://www.crossref.org/working-groups/distributed-usage-logging/

[6] http://www.lagotto.io/

[7] European Commission (2017). Next-generation metrics: Responsible metrics and evaluation for open science. Retrieved from: https://ec.europa.eu/research/openscience/pdf/report.pdf.

[8] Lin, J. & Fenner, M. (2013). Altmetrics in evolution: Defining and redefining the ontology of article–level metrics. *Information Standards Quarterly*, 25(2). Retrieved from: https://www.niso.org/sites/default/files/stories/2017-08/IP_Lin_Fenner_PLOS_altmetrics_isqv25no2.pdf.

many types of potential users who may read and engage with papers without the intention to cite them in a research article, usage statistics offer a promising alternative view from which to judge the impact of open scholarship.

Who Benefits? Values for Stakeholders

There are at least eight primary stakeholder groups which stand to gain added value from newly emerging metrics. The National Information Standards Organization (NISO) completed an *Alternative Assessment Metrics Project* in 2016[9] and formed several working groups aimed at exploring the value of alternative metrics in which they identified the following eight stakeholder groups: librarians, research administrators, hiring committees, funding agencies, academic researchers, publisher/editors, media officers, and platform providers. Use cases explaining how each stakeholder group could benefit from the alternative metrics were formulated, and then sub-divided into three broad themes, which included the ability to: showcase achievements, perform research evaluation, and improve discovery.[10]

Our group also found that usage statistics have been framed as a powerful aid to tracking and understanding investments, as a way to facilitate the advancement of open science, and as a tool that fits within research funding councils' expressed need for open, interoperable, and standards-based alternative metrics.

Usage Statistics to Help Understand Investments

Universities and funding agencies continue to seek new methods of determining return on investment. Any metric that can contribute to understanding performance, at the individual and institutional level, is of use, especially if such a metric is expanded beyond traditional journal metrics.[11]

There is an emergent need within open science/scholarship for metrics that attend to not just the *supply* side of the research picture, but also the *demand* side.[12] Usage statistics represent a *reader-based metric* as opposed to an author-based metric like the citation factor, and therefore provide different insights into research impact and divergent paths of analysis from traditional metric tools. For instance, libraries have traditionally found usage statistics to be of great value in informing acquisitions

[9] National Information Standards Organization (2016). Outputs of the NISO alternative assessment metrics project. NISO RP-25-2016. Retrieved from: https://www.niso.org/publications/rp-25-2016-altmetrics.

[10] Ibid

[11] 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) https://doi.org/10.1045/november2006-organ.

[12] European Commission … Ibid

decisions,[13] which might be in part because of their ability to demonstrate usage and demand that is specific to their local communities.

Demand for openly funded research outputs may also be demonstrable using usage statistics when the statistics are standardized and aggregated across many open repository providers.

Usage Statistics to Advance Openness

In considering metrics' role in supporting and stimulating open science, the European Commission has also suggested that metrics can serve two primary purposes in advancing the support of open science.[14] These purposes are:

- Monitoring the development of the scientific system towards openness at all levels;
- Measuring performance in order to reward improved ways of working at group and individual level.[15]

In order to meet these goals, it is strongly encouraged that new indicators be developed and used responsibly, something which is outlined in documents such as the Leiden Manifesto,[16] the Metric Tide report,[17] and the San Francisco Declaration on Research Assessment.[18] All of these influential reports and initiatives advocate for development and deployment of standards, transparency and openness, interoperability, and for the responsible use of metrics.

There are also incentives for authors that may encourage further participation in openness. The final report of the JISC Usage Statistics Review argues that the value of repository usage statistics resides in their timeliness. Whereas traditional citation metrics take time to aggregate, usage data from repositories are presented more or less instantly, affording authors a more immediate assessment on the visibility of their work.[19] As such, repository metrics function as a recruitment tool. Bruns and Inefuku

[13] Glänzel, W., & Gorraiz, J. (2015). Usage metrics versus altmetrics: Confusing terminology? *Scientometrics*, *102*(3), 2161–2164. https://doi.org/10.1007/s11192-014-1472-7.

[14] European Commission ... Ibid

[15] Performance measurement in this context does not refer to researcher output, but rather to ways of measuring open practices that are not accounted for or acknowledged in traditional reward structures.

[16] Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature News, 520*(7548), 429. https://doi.org/10.1038/520429a.

[17] Wilsdon, J., L. Allen, E. Belfiore, & R. Kain. (2015). The Metric Tide: Report of the independent review of the role of metrics in research assessment and management. https://doi.org/10.13140/RG.2.1.4929.1363.

[18] DORA (2012). San Francisco Declaration on Research Assessment. Retrieved from: https://sfdora.org/read/.

[19] Joint Information Systems Committee (2008). Final report: JISC usage statistics review. Retrieved from: https://repository.jisc.ac.uk/250/1/Usage_Statistics_Review_Final_report.pdf.

(2015) argue that the "collection and reporting of metrics are valuable tools repository managers can exploit to sustain and encourage faculty participation in repositories."[20]

The Need for Metrics Based on Interoperable, Open Standards

There is a multiplicity of alternative metrics to consider as well as a need to make informed decisions as to how to ensure that these are well generated, curated, and used appropriately. The Higher Education Funding Council of England commissioned an Independent Review of the Role of Metrics in Research Assessment and Management in 2015 that summarizes the trend as follows: "There are powerful currents whipping up the metric tide. These include growing pressures for audit and evaluation of public spending on higher education and research." Also noted in the report is that research administrators are now clearly expressing a need for indicators, "underpinned by an open and interoperable data infrastructure."[21]

There is a significant opportunity for repositories to be involved in this *metric tide* if we can ensure that our metrics are useful and that they meet quality assurance standards.

The above rationales for collecting usage data are more easily supported and of greater utility if the statistics are comparable, authoritative, and trustworthy. If this is the case, comparison with publishers is feasible and cost-effective.[22]

The usage statistics being gathered and presented within our repositories simply do not meet this standard at current date. Emerging tools and approaches that are explored within this report will help us take the next steps needed to meet these requirements.

Process

Starting in the Spring of 2018, we undertook two streams of activities.

The first activity was to develop and distribute a set of questions directed at Canadian repository managers. The goal in distributing this survey was to aid our group in understanding the existing practices for collecting and measuring usage statistics within a Canadian context, and to gain insight into the expressed goals and

[20] Bruns, T. & Inefuku, H. W. (2015). Purposeful metrics: Matching institutional repository metrics to purpose and audience. *Digital Scholarship and Initiatives Publications, 4*. Retrieved from: https://lib.dr.iastate.edu/digirep_pubs/4.

[21] Wilsdon, J., Allen, L., Belfiore, E., & Kain, R. (2015) … Ibid

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

aspirations associated with the collection of usage statistics. Findings for this activity are provided in the Responses on IR Usage Statistics Survey section.

In the second activity, we conducted reviews of emerging tools such as IRUS-UK, OpenAIRE Statistics service, and the Repository Analytics & Metrics Portal (RAMP). In tandem, we reviewed the COUNTER Code of Practice Release 5, which includes practices and recommendations for blocking bots, removing inaccurate usage data, and presenting usage statistics in a reliable and comparable format. Short summaries of the services, and concepts we reviewed are available in the Standards and Approaches to IR Usage Data section of this report.

To create these recommendations, we explored the underlying theories, best practices, and known issues that accompany the collection of usage statistics for repositories. Insights gained from this process are reflected throughout this report and are actualized in our suggested recommendations.

Responses on IR Usage Statistics Survey

Method

To understand the landscape and current practice of CARL institutions with respect to Repository statistics and analytics, a survey was conducted. Survey questions are available as part of the Appendix of this document. This survey consisted of ten questions and was presented in both official languages. A single response from each of the CARL institutions was sought. At the time the following statistics were prepared, a total of 42 responses had been received, which constituted a completion rate of 70%. In general, responses could be categorized along thematic lines.

Summary

General Collection and Use Patterns

Summarized results are presented in Figure 1. In general, we observed:

- 2.5% of respondents (n=1) surveyed reported that they did not track repository usage.
- 75% (n=30) of respondents indicated that they shared statistics with the wider campus community.
- 47.5% (n=19) of respondents indicated that they collected and used log files as a source of usage statistics data.

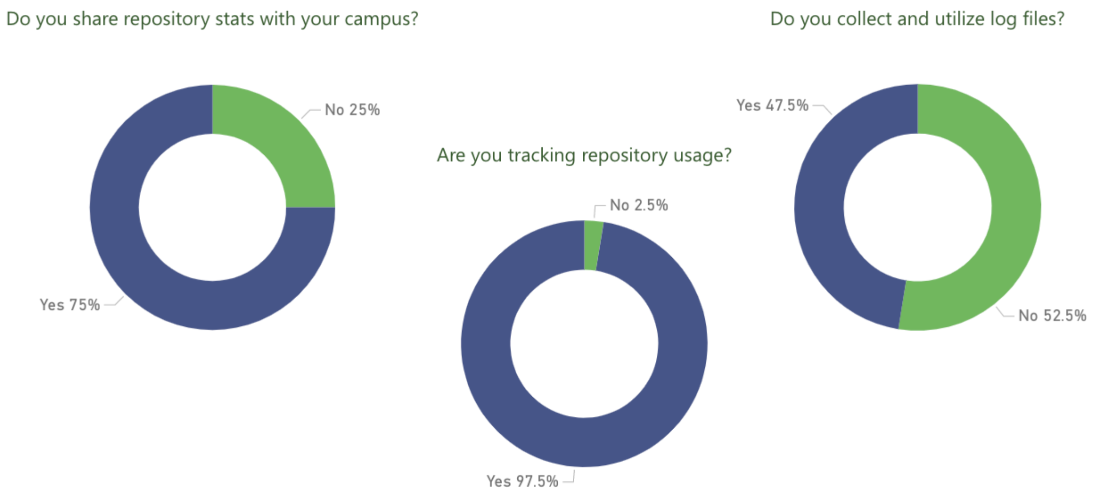

Figure 1.
Tracking and Sharing in Canadian Institutional Repositories

Use of Google Services

Google provides a large suite of freely available tools and methods that web content creators can use to perform modest search engine optimization, and usage analysis. Respondents' usage of Google services are summarized in Figure 2. Without much surprise, the majority of institutions, 85% (n=34) utilize Google Analytics to generate reports to access real-time usage information. Of lesser popularity are two other methods: Sitemap submissions to Google, where only 37.5% (n=15) take advantage, and implementing Google Search Console, used by 45% (n=18). The lower utilization rate of these last two services might speak to their lack of popularity and familiarity for repository administrators. Since cost is not a factor, as all of these services are free, it is reasonable to suggest that these services might be utilized more widely if they were widely promoted amongst repository administrators.

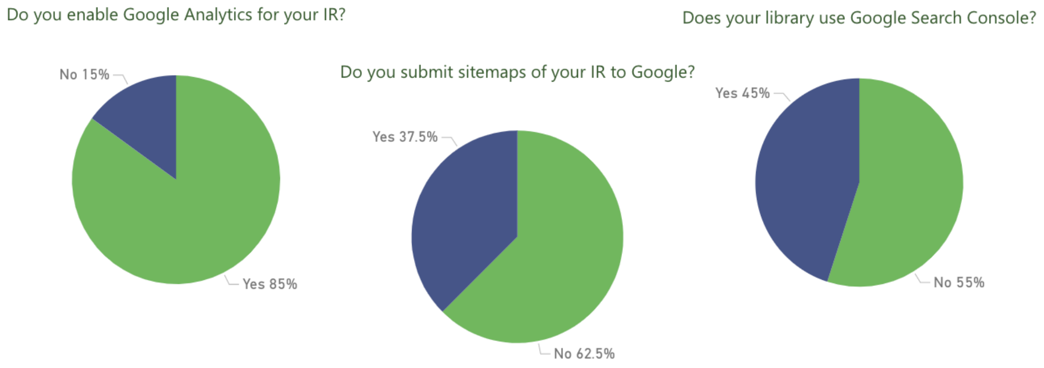

Figure 2.
Google Service Usage in Institutional Repositories

Methods of Usage and sources of Usage Information

Respondents were also asked to describe the various methods by which usage statistics were captured and if automated bot traffic was filtered out. Figure 3 summarizes the responses. Analytics that are classified as internal refer to the built-in analytics systems provided native to the repository platform. External systems refer to any analytics provided by Google, or by a secondary analysis of log files, or via other statistical products outlined in this report. 52% (n=21) of repositories make use of both internal and external analysis tools. An almost complete split down the middle is seen with the filtering of bot (or automated) traffic. Only 47.5% (n=19) actively remove this potential source of noise.

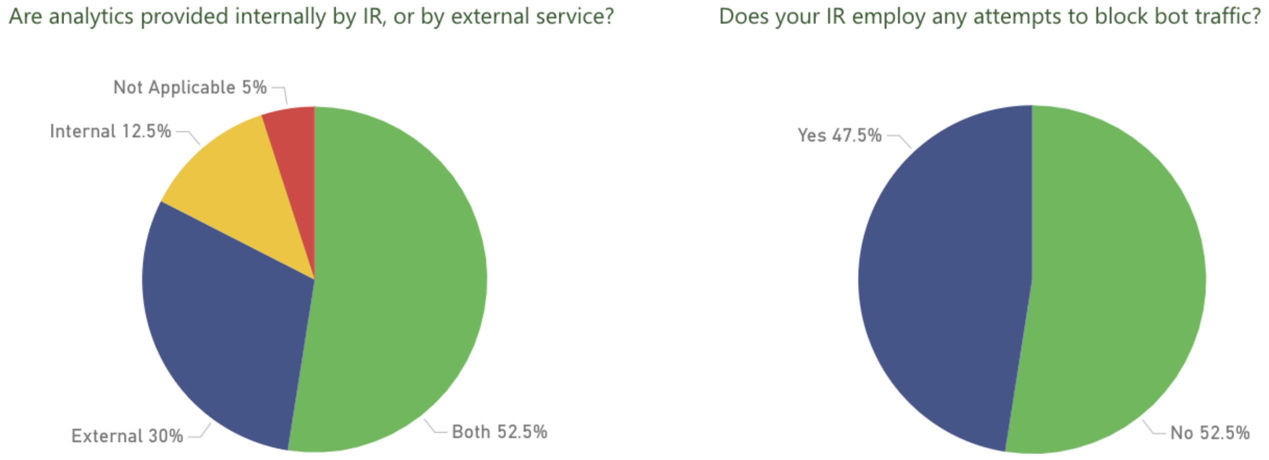

Figure 3.
Technical Aspects of Institutional Repositories

Satisfaction of usage information provided

Survey respondents were asked to comment upon the perceived utility of usage statistics that were shared with stakeholders. Answers varied, but the majority of respondents were at least satisfied with repository statistics services. However, a quarter of respondents chose not to provide an answer, which could indicate that there was not enough experience with this aspect of stakeholder engagement to comment. Figure 4 presents the ratio of responses with respect to this question. 57.5% (n=23) of responses indicated a minimum level of satisfaction from external users with respect to the usage statistics provided, clearly indicating that there is room for improvement with such services.

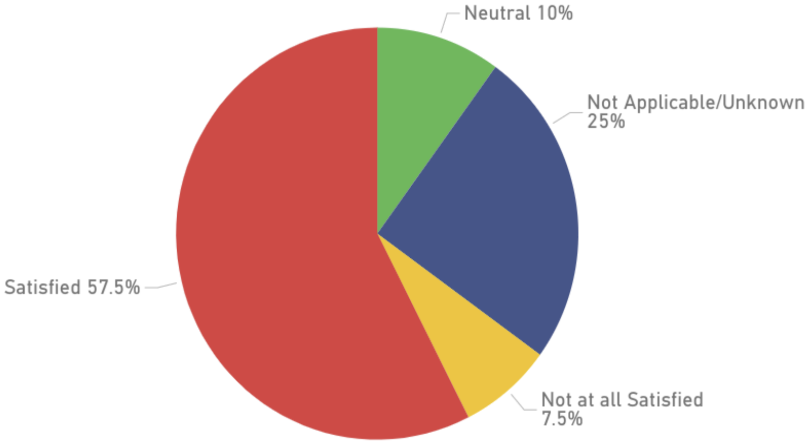

Figure 4.
What is your perception of your users' satisfaction with the statistics you provide?

Standards and Approaches to IR Usage Data

RAMP

The Repository Analytics & Metrics Portal (RAMP) is a web service formed out of a partnership between Montana State University, the Association of Research Libraries, the University of New Mexico, and OCLC Research.

The RAMP team has generated a significant body of research that explains and verifies suspected undercounting of download events when using Google analytics, which is known to only count download events triggered by activity stemming directly from a repository's website. RAMP counts external referrals (direct downloads) from Google searches by incorporating data collected from another Google product called Google Search Console (GSC). GSC data with combined and de-duplicated data from Google Analytics is said to offer a more accurate, and likely increased, count of item usage.

RAMP's approach provides a meaningful distinction between ancillary pages, summary pages, & Citable Content Downloads (CCD). The CCD is posited by RAMP to be the most valuable indicator of scholarly use, due to its being associated with direct access to the scholarly object.

RAMP does not employ its own tactics to block bots, but instead suggests that Google is "one of the best in the world at bot detection" and further justify this choice with the presumption that, "market conditions provide Google with the incentive and resources to invest in bot detection that far exceeds the abilities of the library community."[23]

IRUS UK (Institutional Repository Usage Statistics UK)

IRUS UK is a service to Jisc (Joint Information Systems Committee) members as part of a Jisc subscription. IRUS UK is funded by Jisc and is currently in use in at least 144 institutions.[24]

IRUS UK provides COUNTER-conformant usage statistics for material downloaded from participating institutional repositories via a tracker-code plug-in, which will soon be revised to accommodate COUNTER Release 5. A major benefit of the service is that it provides comparable, consistent, comprehensive, and standards-based metrics that help to determine the value and impact of institutional repositories.[25]

Institutions outside of the UK may use IRUS UK as a hosted service at a cost. Data is transmitted to OpenAIRE for any repository using IRUS, relieving those IRs of having to manage OpenAIRE compliance. OpenAIRE integration would be possible for any Canadian institution cooperating with IRUS UK. *The International collaboration and value: working with OpenAIRE* case study provides more details on OpenAIRE integration.[26]

IRUS UK is now open to institutions outside of the UK and a pilot US project occurred in 2018. OAPEN, the University of Amsterdam, and various repositories in Australia and the United States are currently using the service. IRUS is currently discussing further options with regard to providing the service in North America.

Patches are currently available for DSpace. Other options include plug-ins for EPrints and Haplo, and a Ruby Gem for Hydra and Samvera repositories. Pure and Worktribe have implemented tracker functionality for their repository software platforms. IRUS has also had conversations with other vendors, such as Bepress.[27]

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

[24] As reported by OpenAIRE (n.d.), National Open Access Desk, Source: https://www.openaire.eu/item/united-kingdom.

[25] IRUS-UK (n.d.). Frequently asked questions. Retrieved from: http://irus.mimas.ac.uk/support/faqs/.

[26] IRUS-UK (2018). International collaboration and value: Working with OpenAIRE case study. Retrieved from: https://irus.jisc.ac.uk/documents/IRUS-UK_working_with_OpenAIRE.pdf.

[27] See: http://irus.mimas.ac.uk/ for guides, toolbox, support, case studies, use cases, FAQs, and tips.

As of 2018, Hyrax does not include an admin UI for auditing Fedora's built-in fixity checks, or for restoring objects from backups if bitrot or file corruption is detected[28].

OpenAIRE Statistics

OpenAIRE is a European-based organization that seeks to shift scholarly communication towards open through a variety of different initiatives which include policy creation, infrastructure development, training, advocacy, standard development and, more specifically, services for interoperability of repositories and repository usage.[29]

More specifically, OpenAIRE provides the OpenAIRE Usage Statistics Service,[30] a comprehensive dashboard of analytics of repository usage that is powered by the Matomo web analytics platform.[31]

The statistics themselves are produced using the COUNTER Code of practice directives. Much like most of OpenAIRE's offerings, the usage statistics service is fairly comprehensive. For example, it can perform deduplication of the same item in multiple repositories so that reporting aggregates and merges accordingly.

Project COUNTER

Project COUNTER, founded in 2003, is an international non-profit organization that develops a Code of Practice standard for usage data counts. This standard is a well-known and reliable protocol, currently in use by many libraries and vendors internationally. Employing the ANSI/NISO Z39.930-2014 protocol, it counts usage data for the following electronic resources in libraries: journals, databases, datasets, books, book segments, reference works, multimedia databases, newspapers, repository items, reports, and theses or dissertations.

The standard attempts to provide "consistent, credible and comparable usage data."[32] This service assists librarians in assessing the value of resources and helps publishers support that goal by providing statistics comparable across various library vendors. The process is supported by two approved COUNTER auditors. Compliance with Release 5, which aims to reduce the complexity of the code, address evolving needs,

[28] Rochkind, J. (2017). Exploring and planning with Sufia/Hyrax/Fedora fixity validation. https://bibwild.wordpress.com/2017/05/01/exploring-and-planning-with-sufiahyraxfedora-fixity-validation/.

[29] OpenAIRE (n.d.). Mission and vision. (website). Retrieved from: https://www.openaire.eu/mission-and-vision.

[30] http://catalogue.openaire.eu/service/openaire.openaire_usage_statistics

[31] https://matomo.org/

[32] COUNTER (n.d.) Landing page (website). https://www.projectcounter.org/.

provide greater customization, and simplify maintenance, was required by January 2019.

In addition to usage statistics, the COUNTER Code of Practice framework defines the data elements to be measured, the definitions of these elements, usage reports, specifications for data processing, requirements for the auditing process, and guidelines to avoid duplicate counting.

COUNTER is augmented by SUSHI (Standardized Usage Statistics Harvesting Initiative), a tool that facilitates the retrieval of COUNTER usage data, thereby eliminating the need to retrieve data from each web site independently, and the COUNTER Report Validation Tool, which permits vendors and libraries to test the implementation of SUSHI and COUNTER reports.

Some statistical services under review in this document are founded on the COUNTER protocol, which confers a significant degree of confidence in COUNTER's data-collection methodology and reporting.

Google Analytics

Google Analytics, launched in late 2005, is a web analytics service that collects, measures, and analyzes web traffic through the use of a tracking code. In addition to assisting with understanding return on investment, it is a market research tool and a means to optimize website effectiveness.

Google Analytics focuses on Dimensions and Metrics. Dimensions presents data attributes, such as geographic origin of traffic and the page being measured. Metrics are quantitative measures of that data, such as bounce rate, session duration, pageviews per session, average time on page, percent exit, etc. Data is categorized as follows: acquisition, meaning how one gets website traffic; behaviour, meaning what visitors are doing on the website; and conversions, meaning a completed activity.

Several issues may interfere with the correct collection of data. The most common of these are browsers with JavaScript disabled, users who refuse cookies, the same user using two different devices, as well as ad filtering, or the use of private networks, both of which prevent some data from being collected. Google Analytics does not present real-time data and may include referral spam. In addition, Google may employ data sampling for very active websites.

Research has suggested that Google Analytics is inappropriate for institutional repositories because it "fails to capture the vast majority of non-HTML Citable Content

Downloads" from repositories. Despite this fact, the tool is in use in a large number of repositories.[33]

Log File Analysis

A key characteristic of systems that function as web servers is that they generate a file containing all of the transactions that they perform. This file is written in plain text and is referred to as a log file. This file is structured according to a standardized format[34] that is known as the common log format (CLF). Each line in a CLF file is tab delimited and is comprised of the essential pieces of each action performed by the server. These components include IP address, date/time stamp, the file served in the request, status code, byte size of file, plus others.

Since these log files describe every action the server performs, they can often grow very large and can quickly get into the gigabyte range. On account of how verbose they can be, examining them directly is tedious. As a response, many software packages have been created to handle aggregating and presenting trends found in these files. Popular platforms include: Analog[35], and Graylog[36]. A large portion of these platforms are free and open source, although many commercial solutions exist. The types of analysis they can provide is expansive but includes things such as geographic location of visitors (based on IP address), identifying popular content, and providing insights on any errors that might be occurring, such as users requesting content that does not exist (e.g. 404 errors).

Since repositories are essentially web servers for a specific type of content, they generate CLF files that can provide insights about usage. There are some challenges, however, with accessing the files themselves that might prove to be detrimental. For example, with a cloud-based repository platform, the hosting provider might not have a mechanism to expose log files for analysis. Additionally, running the software that performs the analysis places an additional burden on the Library, which now needs to run an additional platform.

ROBOTS (Bots)

The tendency for some web traffic to be initiated by computer programs, also referred to as bots, can make it difficult to distinguish or differentiate real human usage, and thus creates uncertainty regarding the accuracy of reported statistics.

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

[34] IBM. Log file formats: NCSA Common. http://publib.boulder.ibm.com/tividd/td/ITWSA/ITWSA_info45/en_US/HTML/guide/c-logs.html#common.

[35] Analog CE. https://www.c-amie.co.uk/software/analog/

[36] Graylog. https://www.graylog.org/

“Robots outnumber humans 10:1 in terms of sessions, 5:4 in terms of raw HTTP accesses, and 4:1 in terms of megabytes transferred”.[37]

The accepted solution within most web services is to require a username and password in order to access the content.[38] Open collections that are intended to be free from access barriers do not have this option, and instead are met with the cumbersome task of either blocking bots proactively or sorting through results to remove bot traffic retroactively.

COUNTER provides a list of well-known robots, whose usage should be removed in order to meet the requirements of their audit process. The list is, “not intended to be a comprehensive list. The need for more sophisticated rules and processes is well understood”.[39] There are also community-maintained lists of known bots and spiders which can be added to a file on the web server called robots.txt. In addition, there are community-maintained lists of malicious IP addresses maintained by Project Honeypot, which can be used to train and build better filters.

For services that utilize Google Analytics, there is an administrative option that is not turned on by default but can be enabled to ‘exclude all hits from known bots and spiders. The bot filtering option in Google Analytics utilizes a paid list of bot filters that are provided by the Interactive Advertising Bureau (IAB), which are regularly updated and maintained.[40]

There are certain ‘good bots’ that identify themselves plainly, such as Google bots that crawl websites to create search indexes. However, filtering bots proactively is problematic because some bots are programmed deceptively to behave like regular users. One approach uses adaptive filtering, which aims to progressively improve bot detection, and filtering using algorithms. However, “the more refined the filtering system the more likely that regular users will be excluded along with the unwelcome ones”,[41] which is not ideal for open resources. Greene (2016) conducted a

[37] AlNoamany, Y. A., Weigle, M. C., & Nelson, M. L. (2013). Access patterns for robots and humans in web archives. *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, 339–348. https://doi.org/10.1145/2467696.2467722.

[38] Amshey, S. (n.d.) Bot shields: Activate! ensuring reliable repository download statistics. BePress. Webinar. Retrieved from: https://www.bepress.com/webinar/bot-shields-activate-ensuring-reliable-repository-download-statistics/.

[39] IRUS-UK (2013). Position statement on the treatment of robots and unusual usage. Retrieved from: https://irus.jisc.ac.uk/documents/IRUS-UK_position_statement_robots_and_unusual_usage_v1_0_Nov_2013.pdf.

[40] Moore, A. (2015). Eliminating bot traffic from Google Analytics once and for all. Retrieved from: https://www.bounteous.com/insights/2015/04/01/eliminating-bot-traffic-google-analytics-once-and-all/?lang=en-ca.

[41] Information Power LTD. (2013). IRUS download data – identifying unusual usage. Retrieved from: https://irus.jisc.ac.uk/documents/IRUS_download_data_Final_report.pdf.

comprehensive review of 10 different approaches to robot detection concluding that, “no method is capable of accurately detecting all robots visiting a given web server” and instead suggests that the goal of these techniques is in “capturing as many robots as possible while labelling the fewest number of human sessions as robots”.[42]

The use of filters for open scholarly communications platforms is in its early days of experimentation and understanding. Prior to the latest release of the COUNTER Code of Practice, COUNTER formed a ROBOTS Working Group that was tasked with investigating the possible use of filters. They experimented with several techniques and describe their work as “a beginning of a set of standards to be built upon that will help in achieving consistent, credible and comparable usage statistics that can be aggregated across many types of scholarly communication platforms.”[43]

A web standard called the robots exclusion protocol is designed to inform robots of how to interact with a website by flagging pages that are to be excluded from crawling. Not all robots comply with the web standard, and some will ignore it, so this method is not sufficient on its own. The exclusion list provided by the robots.txt file can be complemented by sitemaps, which provide a detailed list of the site’s structure, along with flagged areas of the site which can be included in web crawls. Site maps help with discovery and indexing of a site’s contents and having a well-structured, easy to navigate web page is a known way to increase the likelihood of a high search engine results page (SERP) rank.[44] This suggests another reason why blocking bots proactively might be disadvantageous to repository managers’ goals of having their IR’s content be discoverable.

Important questions also arise as to whether blocking patterns of automated usage may exclude legitimate usage; “automated downloads are not necessarily robots, e.g. an institution doing a major literature search uses a script to cross search a number of databases and repositories”.[45] Attempts to address this include COUNTER’s addition of the Access Method “TDM” attribute, which is designed for report creators to flag instances of Text & Data Mining.[46]

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

[43] Greene, J. W. (2017). Developing COUNTER standards to measure the use of Open Access resources. *Qualitative and Quantitative Methods in Libraries, 6*(2), 315-320. Retrieved from: http://www.qqml-journal.net/index.php/qqml/article/view/410.

[44] Arlitsch, K., OBrien, P., & Rossmann, B. (2013) Managing search engine optimization: An introduction for library administrators (preprint), *Journal of Library Administration, 53*(2-3), 177-188. Retrieved from: https://scholarworks.montana.edu/xmlui/handle/1/8671.

[45] IRUS-UK (2013). Position statement … Ibid

[46] Mellins-Cohen, (n.d.). The friendly guide to release 5: Technical notes for providers. Retrieved from: https://www.projectcounter.org/wp-content/uploads/2017/07/Tech_Notes_20170710.pdf.

Regardless of these complicated issues, bots are responsible for almost 50 percent of internet traffic and 85 percent of IR downloads,[47] and it is in our best interests to continue exploring ways to better differentiate and define genuine usage patterns, in order to more accurately reflect the value of our repositories.

Concluding Remarks

This report has outlined the important role that IR statistics can play in research assessment, but more so, in the advancement and understanding of our ongoing investments into open scholarship.

There does appear to be a modest to high level of effort and interest being displayed across the surveyed Canadian institutions to track repository usage, and it is clear that Google's products, along with the 'out of box' statistical tools provided by IR platforms, are the prominent tools of choice.

Our exploration of the emerging tools and approaches for improvement and standardization of IR usage statistics revealed that this is a rich and burgeoning field of academic and technical study that goes well beyond what an 'out of box' approach to statistics can provide. The need for open and interoperable standards is evident, as is the call to protect the privacy of users when gathering statistics.

Regarding consistency and reliability, COUNTER provides the only internationally recognized standard for collecting and reporting statistics about scholarly materials being disseminated on the web. COUNTER has also been forming working groups to tackle the complex problems associated with Robot usage, has implemented API reporting through creation and maintenance of the SUSHI protocol, and has released a code of practice for applying the same principles to research data.

New services acting as central clearing houses for IR statistics are offering tools and expertise that perform necessary tasks of standardization by utilizing the COUNTER code of practice. This offers Canadian IRs a chance to greatly improve the reliability and consistency of their statistics.

Our group found that both IRUS-UK and OpenAIRE Statistics offer the most favourable approaches to the collection and presentation of IR Usage Statistics due to their focus on generating COUNTER compliant statistics. RAMP has not adopted a recognized, standardized approach to the collection and reporting of IR Statistics. However, we applaud RAMP for their ongoing efforts to create a unique and easy to use dashboard service. Their focus on surfacing external referrals to PDF views is

[47] OBrien, P., Arlitsch, K. Mixter, J., Wheeler, J. & Sterman, L. (2017) "RAMP – the Repository Analytics and Metrics Portal" … Ibid.

novel and worth studying. Although IRUS-UK and OpenAIRE deploy different approaches, they're quite similarly capable of meeting the requirements that we've set out. IRUS-UK represents a worthy alternative that would be equally deserving of consideration as a standardized service to all Canadian IRs.

In conclusion, we recommend that Canadian IRs choose to collectively adopt OpenAIRE Statistics which is ideal due to its use of the privacy-focused Matomo web analytics tool, and due to its global approach to working with open repositories. This choice also aligns with other work of the Open Repositories Working Group, which continues its work to integrate with international initiatives for the dissemination of open scholarship.

Appendix 1: Survey Questions

Name of your Institution || Nom du dépôt

What is your Repository's URL? || Quel est l'URL de votre dépôt?

Are you Tracking Repository Usage? || Suivez-vous l'utilisation de votre dépôt?

Are analytics provided internally by IR, or by external service? || Est-ce que l'analyse (« analytics ») est comprise dans le dépôt, ou utilisez-vous un service externe?

Do you enable Google Analytics for your IR? || Vous servez-vous de Google Analytics pour votre dépôt?

Does your library use Google Search Console? || Est-ce que votre bibliothèque utilise Google Search Console?

Do you submit sitemaps of your IR to Google? || Contribuez-vous les plans de site de votre dépôt à Google?

Does your IR employ any attempts to block bot traffic? || Est-ce votre dépôt emploie des moyens pour bloquer les robots (« bots »)?

What would you like to know about your IR's usage? || Qu'est-ce que vous désirez savoir au sujet de l'utilisation de votre dépôt?

Do you collect and utilize log files? || Est-ce que vous recueillez et utilisez les fichiers journaux (« log files »)?

Do you share repository statistics with your campus? || Diffusez-vous les statistiques de votre dépôt avec les membres de votre campus?

What is your perception of your user's satisfaction with the statistics you provide? || D'après vous, quel est le niveau de satisfaction de vos usagés quant aux statistiques que vous diffusez?