Received 4 August 2020
Revised 10 September 2020
Accepted 14 September 2020

An analysis of use and performance data aggregated from 35 institutional repositories

Kenning Arlitsch
Library, Montana State University Bozeman, Montana, USA
Jonathan Wheeler
University Libraries, University of New Mexico, Albuquerque, New Mexico, USA
Minh Thi Ngoc Pham
School of Information Science and Learning Technologies, University of Missouri, Columbia, Missouri, USA, and
Nikolaus Nova Parulian
School of Information Sciences, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA

Abstract

Purpose – This study demonstrates that aggregated data from the Repository Analytics and Metrics Portal (RAMP) have significant potential to analyze visibility and use of institutional repositories (IR) as well as potential factors affecting their use, including repository size, platform, content, device and global location. The RAMP dataset is unique and public.

Design/methodology/approach – The webometrics methodology was followed to aggregate and analyze use and performance data from 35 institutional repositories in seven countries that were registered with the RAMP for a five-month period in 2019. The RAMP aggregates Google Search Console (GSC) data to show IR items that surfaced in search results from all Google properties.

Findings – The analyses demonstrate large performance variances across IR as well as low overall use. The findings also show that device use affects search behavior, that different content types such as electronic thesis and dissertation (ETD) may affect use and that searches originating in the Global South show much higher use of mobile devices than in the Global North.

Research limitations/implications – The RAMP relies on GSC as its sole data source, resulting in somewhat conservative overall numbers. However, the data are also expected to be as robot free as can be hoped.

Originality/value – This may be the first analysis of aggregate use and performance data derived from a global set of IR, using an openly published dataset. RAMP data offer significant research potential with regard to quantifying and characterizing variances in the discoverability and use of IR content.

Peer review – The peer review history for this article is available at: https://publons.com/publon/10.1108/OIR-08-2020-0328

Keywords Webometrics, Digital repositories, Institutional repositories, IR, Academic publishing, Web analytics

Paper type Research paper

The authors are grateful to the managers of the 35 repositories described in this study, who recognized the value of participating in RAMP.

This research was conducted as part of a grant generously funded by the Institute of Museum and Library Services (IMLS Log Number: LG-72-18-0179).

Online Information Review
Vol. 45 No. 2, 2021
pp. 316-335
Emerald Publishing Limited
1468-4527
DOI 10.1108/OIR-08-2020-0328

Introduction

The Repository Analytics and Metrics Portal (RAMP) is a web service that aggregates performance and use data of institutional repositories (IR); it produces a unique dataset of standardized metrics across time [1]. Developed by Montana State University, the University of New Mexico, OCLC Research and the Association of Research Libraries (OBrien *et al.*, 2017), RAMP currently aggregates data from more than 55 repositories around the world, and new repositories continue to be added. RAMP tracks items from registered repositories that have surfaced in search results across all Google properties, including data that show whether users clicked through to the IR and downloaded the file. This article demonstrates basic use and performance data for 35 repositories in seven countries that were registered with the RAMP from January 1–May 31, 2019.

The RAMP dataset is composed of Google Search Console (GSC) data aggregated from registered repositories. It is the first openly published dataset of use and performance data aggregated from a cross platform set of IR in multiple countries, using a common set of metrics. GSC provides accurate non-HTML download counts executed directly from all Google search engine results pages (SERP). Metrics include impressions (the number of times an item appears in the SERP); position (the location of the item in the SERP); clicks; click-through ratios; date; device and country (Frost, 2019; Google, Inc., 2020). RAMP data are collected from GSC in two separate sets: page-click and country-device data. The page-click data include the Uniform Resource Locator (URL) of every item that appeared in the SERP, which opens significant possibilities for additional research if the metadata of those items were mined.

The basic analyses of RAMP data performed for this paper demonstrate large performance variances across IR as well as low overall use. The data confirm that device use affects search behavior, that different content types such as electronic theses and dissertations (ETD) may affect use and that searches originating in the Global South show much higher use of mobile devices than in the Global North. In addition to providing a baseline for IR performance metrics, the data offer significant research potential. RAMP data can help scholars understand the disciplinary scope and breadth of the content contained in the IR.

Research questions

The research questions are designed to elicit a data-driven story about IR that has not previously been available due to a lack of a uniform analytics model applied across a disparate set of repositories.

(1) What do RAMP data show about the size and the use of IR?

(2) How do devices used to access IR content affect search behavior?

Literature review

Calls for standardized reporting for IR have been evident almost as long as they have existed (Fralinger and Bull, 2013; Harnad and McGovern, 2009; Organ, 2006). However, nearly all such calls end with some variation of the lament that “no clear standard has yet emerged for repository reporting and assessment” (McDonald and Thomas, 2008). The siloed and distributed nature of the IR landscape has exacerbated this situation as research libraries collectively commit significant resources toward running their own IR (Arlitsch and Grant, 2018) and make “long-term commitment[s] for safeguarding, preserving and making accessible the intellectual content of an institution” (Lagzian *et al.*, 2015) despite having little idea of how much they are really being used and with no way to compare against other institutions. Hosted solutions like BePress's Digital Commons platform have the capacity to collect aggregate data, but due to the commercial orientation of this platform these data are not openly available to the IR community.

Bruns and Inefuka assert that metrics must be contextualized to be useful and that download statistics are only part of assessing IR performance, but when discussing the collection download counts from the big three IR platforms (DSpace, EPrints and Digital Commons), they suggest using platform-generated statistics and do not question the accuracy of those numbers (Bruns and Inefuku, 2015). Some research has indicated that platform-generated statistics may be inflated by as much as 85% because they rely on log file analysis, and nonhuman traffic is very difficult to filter (Greene, 2016). Other researchers also correctly warn of the limitations associated with placing too much faith in statistics to determine quality, use and performance of repositories. They point out that it is difficult to distinguish different kinds of traffic, that items may have multiple URLs and that different kinds of collections act differently with search engines, among other limitations (Perrin *et al.*, 2017). The importance of standardized usage statistics "across repositories and repository platforms" is highlighted in the Next Generation Repositories Report (Rodrigues *et al.*, 2017).

The IRUS-UK project is similar in intent to RAMP and has been operating since 2012 (Needham and Stone, 2012). IRUS-UK collects data from approximately 180 repositories in the United Kingdom (Jisc, 2019), and a pilot project known as IRUS-USA was launched to test the IRUS model with 11 repositories in the United States (Kim, 2018). An assessment of the IRUS-USA pilot project concluded that the IRUS model was favorably received, and that participants desired more documentation, more functionality and granularity available in the reports and more visual statistics (Thompson *et al.*, 2019).

A principal difference between IRUS and RAMP is the IRUS Tracker Protocol that must be installed at each repository. It "gathers basic raw data for each download," which are then processed to create COUNTER-compliant data and to remove "robot activity or other unusual activity" (MacIntyre and Jones, 2016). IRUS-UK attempts the difficult process of filtering robot activity from log files and then performs various analyses on the cleaned data. Filtering is defined in an IRUS-UK position statement: "COUNTER provides a list of robots, whose usage should be removed as a bare minimum. The list is used as part of the audit process and is not intended to be a comprehensive list. The need for more sophisticated rules and processes is well understood." The document also states that IRUS-UK has "added further filters to remove more user agents identified as robots and applied a simple threshold for 'overactive' IP addresses" (IRUS-UK Team, 2013).

The RAMP requires no local installation, as it simply utilizes the data that Google passes to it through the GSC Application Programming Interface (API) (Google, Inc., 2020). RAMP also relies on Google to filter robot activity. Google's success as an advertising platform depends on its ability to guarantee to customers that clicks are human generated. The data that are passed to RAMP through GSC are therefore as robot free as can be hoped, and the relatively conservative download numbers demonstrated by RAMP, compared to platform-generated statistics, support this theory (OBrien *et al.*, 2017). The RAMP misses non-Google referrals to repositories, but prior work by the research team indicates that a vast majority of traffic to IR is driven by Google properties and that the traffic referred by other search engines, social media sites, email, etc. is comparatively small. Therefore, while the sheer numbers of referrals and downloads tracked by RAMP may be conservative, we believe that the trade-off for nearly robot-free data is worthwhile.

Methodology

The findings reported below represent a webometric analysis of RAMP data. The term “webometrics” was coined to describe research into network-based information using quantitative measures (Almind and Ingwersen, 1997) and includes the application of mathematical and statistical methods to the study of quantitative aspects of information sources on the World Wide Web. Webometrics regards web pages as “information entities,” with hyperlinks across pages acting as citations and citation networks (Almind and Ingwersen, 1997).

Webometrics includes four main areas of research: (1) web page content analysis; (2) web link structure analysis; (3) web usage analysis, which includes analysis of users’ search and browsing behavior and (4) web technology analysis, which includes search engine performance (Björneborn and Ingwersen, 2004). The current study focuses on two areas of webometric analysis: web usage and web technology. Data analytic methods including data aggregation, data reshaping, visualization and descriptive statistics are used to examine the performance of RAMP participants, including the use of IR items, their discoverability in search engines and technologies used to search and view the IR content.

Data creation and analysis methods

The published data include monthly CSV files of page-click and country-device data for the period of January 1–May 31, 2019 (Wheeler *et al.*, 2020b). Python and *R* scripts used for the analysis are available on GitHub (Wheeler *et al.*, 2020a). More detailed versions of the Tableau visualizations shown in this paper may be viewed at Tableau Public (Parulian, 2020).

The number of repositories currently registered with RAMP exceeds 55, but data from only 35 repositories were analyzed for this research (Table 1). The reason for this is consistency. Most of the other repositories were registered more recently and had not accumulated data for the same five-month period of time (January 1, 2019–May 31, 2019). A few had also experienced configuration errors that were usually the result of updates to the repository platform. The authors have a high degree of confidence in the accuracy and consistency of the data from 35 repositories for the period under study.

The dataset created for this study is available for download from the Dryad data repository, along with documentation (Wheeler *et al.*, 2020b). In order to calculate summary statistics relative to overall IR content, additional data were gathered for the page-click analyses described below (Arlitsch *et al.*, 2019). These data include the total number of items hosted by each repository, the corresponding IR platform, the country in which the IR is located and the count of electronic theses and dissertations (ETDs) in each IR.

RAMP data are harvested daily via the GSC API (Wheeler *et al.*, 2020b). Two datasets are downloaded per IR: a page-click dataset that includes granular data per URL; and a country-device dataset that is less granular and does not include URLs. The lack of URLs within the country-device data means that these data cannot be combined or cross referenced with page-click data. Another implication for the results is that statistics derived from country-device data represent the aggregate SERP performance of all pages within an IR, including HTML pages and citable content pages (definition below). It is not possible within country-device datasets to disaggregate data about citable content from other activity.

The analyses are therefore divided into two sections. The first section demonstrates a baseline analysis of repository use and citable content downloads (CCD) using page-click data. Data aggregation scripts and resulting summary statistics per analyzed IR are provided on GitHub (Wheeler *et al.*, 2020a).

Country	Type	Repository name	Platform	#Items
The USA	University	Deep Blue (*U* Michigan)	DSpace	124,436
		Digital Commons @*U* Nebraska Lincoln	Digital Commons	105,065
		Digital Repository (UNM)	Digital Commons	93,564
		Caltech Authors	EPrints 3	81,000
		Caltech THESIS	EPrints 3	9,814
		ScholarWorks (*U* Montana)	Digital Commons	75,715
		VTechWorks	DSpace	72,275
		ScholarWorks (*U* Texas)	DSpace	60,359
		Rucore (Rutgers)	Fedora	43,008
		K-REX (Kansas State University)	DSpace	37,084
		UKnowledge (*U* Kentucky)	Digital Commons	34,196
		Digital Scholarship at UNLV	Digital Commons	23,896
		D-Scholarship@Pitt	EPrints 3	21,358
		DRUM (*U* Maryland)	DSpace	21,246
		IUPUI ScholarWorks	DSpace	15,000
		Montana State ScholarWorks	DSpace	14,381
		Swarthmore Works	Digital Commons	11,095
		Digital Repository Service (Northeastern)	Fedora/Samvera	5,446
		NKU Digital Repository	DSpace	2,420
		Scholarly Works @ SHSU	DSpace	2,095
	Consortium	Mountain Scholar	DSpace	71,708
		ShareOK	DSpace	64,972
		Maryland SOAR	DSpace	9,359
		TriCollege Libraries IR	DSpace	8,728
Australia	University	Research Online (*U* Wollongong)	Digital Commons	69,396
The United Kingdom	University	Strathprints	EPrints 3	51,386
		PEARL (*U* Plymouth)	DSpace	9,903
Canada	University	MacSphere (McMaster)	DSpace	17,979
		UWSpace (*U* Waterloo)	DSpace	13,021
	Consortium	VIURRSpace	DSpace	10,366
Sweden	University	Epsilon Archive for Student Projects	EPrints 3	11,895
		Epsilon Open Archive	EPrints 3	9,307
New Zealand	University	Massey Research Online	DSpace	12,207
South Africa	University	Western Cape Research Repository	DSpace	5,143
		Western Cape ETD Repository	DSpace	3,640

Table 1.
List of 35 IR by country, type, name, platform and number of items

The second section demonstrates an analysis of trends evident in the country-device data. Specifically, we examine the number of visits from each country and characterize some limited effects devices seem to have on user search behaviors.

In both cases, it is important to remember that the datasets represent 35 repositories on four different platforms, and that the numbers shown are not necessarily indicative of characteristics of those platforms. Although the majority of RAMP IR host a range of general-purpose academic content including self-archived copies of scholarly articles, datasets, and ETD, some of the analyzed repositories focus more narrowly on hosting specific kinds of content. Some types of content may drive use more than other types. Consequently, the authors reiterate that the results presented below represent a baseline analysis of a unique, open dataset. The authors recommend further study with a larger sample of IR and a longer date range in order to better generalize results across the IR ecosystem at large.

Data definitions

Combining the descriptive data reported by (Arlitsch *et al.*, 2019) with IR search engine performance data from RAMP allows further analysis of IR characteristics – size, platform and location – in terms of the actual access and use of IR content. The page-click analysis and discussion rely on the following definitions:

(1) *Citable content URLs:* Page-click data harvested from the GSC API include search engine performance statistics for individual URLs. Many of these URLs point to ancillary IR pages such as the IR homepage or the HTML pages that contain item level metadata. RAMP maintains these data but primarily reports click activity on content files (PDF, CSV, etc.), so it is necessary to differentiate URLs that point to content files from those that point to HTML pages. URLs that point to content files are referred here as "citable content URLs."

(2) *Citable content downloads (CCD):* These are RAMP's primary metric and are defined as clicks on citable content URLs.

(3) *Item:* An "item" is any single asset published within an IR, including item level metadata and all content files or bitstreams associated with the asset. Items are commonly represented by HTML pages. For example, everything found at https://scholarworks.montana.edu/xmlui/handle/1/9939; it is considered part of a single "item" (Obrien *et al.*, 2016), which includes the published metadata and the seven associated bitstreams or content files. By contrast, each of the seven content file URLs is considered a single citable content URL.

(4) *Use ratio:* For each IR, the use ratio is the count of unique items with CCD divided by the total count of items hosted by the IR. Since a single item within an IR may contain many content files, this calculation requires inferring the HTML URL of the corresponding IR item page for any citable content URL with a positive click value in the RAMP page-click dataset. The inferred HTML URLs are further processed to deduplicate items that occur in the data with both secure (HTTPS) and non-secure (HTTP) connection protocols. The final count of deduplicated URLs is the numerator of this ratio. Using the example, above, if each of the seven content files accessible from (Obrien *et al.*, 2016) were clicked on during the period of study, all of that activity would be aggregated as one item use under the single, corresponding item HTML URL for the sake of calculating the use ratio. Python code and documentation for inferring item pages and calculating use ratio is available at (Wheeler *et al.*, 2020a).

More information about citable content URLs and CCDs, including how data harvested from the GSC API are processed to identify citable content URLs, is available in the published dataset documentation (Wheeler *et al.*, 2020b).

Further limitations of the dataset are noted in the Discussion section of this paper.

Results of the page-click data analysis

RQ1. What do RAMP data show about the size and use of IR?

The number of items hosted by individual repositories ranges from 2,095 to 124,436 (Figure 1); the average number of items in each repository is 34,928 items ($M = 34{,}928$) (Table 2) and the IR

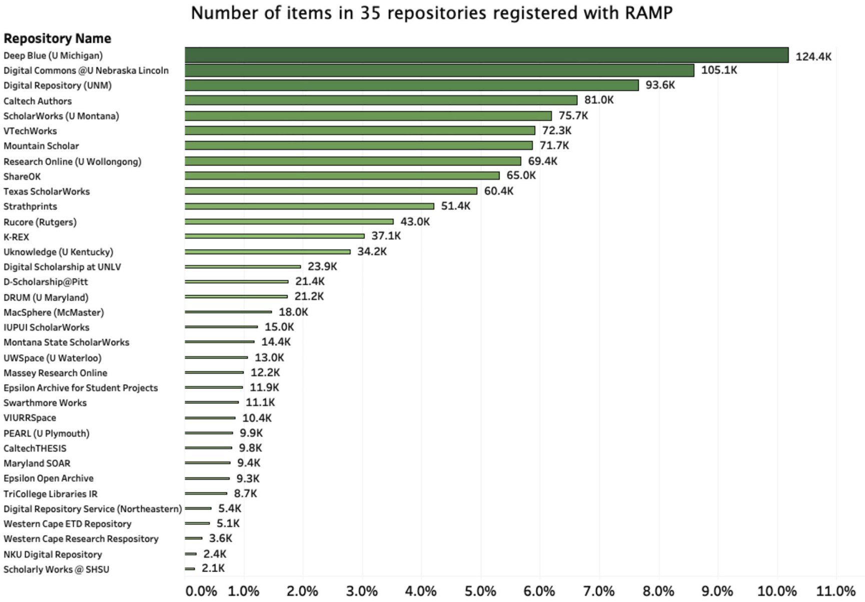

Figure 1.
Number of items in 35 repositories registered with RAMP

Table 2.
Overall statistics of items in repositories registered with the RAMP

Variable	*N*	Min	MDN	*M*	Max
Repository	35	2,095	17,979	34,928	124,436

Note(s): *N* = repositories, *M* = mean and MDN = median

operate on four platforms: Digital Commons, DSpace, EPrints and Fedora (Table 3). The four platforms host 7, 20, 6 and 2 RAMP-registered repositories, respectively.

DSpace is the platform that is used to host the most repositories in the dataset (20) and consequently contains the most items (576,322). Digital Commons is the platform with the second most items (412,927 items) in seven repositories. There are six repositories totaling 184,760 items using the EPrints platform. Fedora has the fewest items with 48,454 items in two repositories. On average, each repository in DSpace has 28,816 items. The EPrints platform ranks second for the average number of items with 30,793. Digital Commons repositories contain an average 58,793 items, almost double the average number of items in DSpace repositories [2].

The repositories are hosted in seven countries: Australia, Canada, New Zealand, South Africa, Sweden, the United Kingdom and the USA (Table 4). Most of the 35 repositories

Table 3.
Number of items by platform

Platform	*N*	Min	MDN	*M*	Max	Total
DSpace	20	2,095	13,701	28,816	124,436	576,322
Digital Commons	7	11,095	69,396	58,990	105,065	412,927
EPrints	6	9,307	16, 627	30,793	81,000	184,760
Fedora	2	5,446	24,227	24,227	43,008	48,454

Note(s): *N* = platform, *M* = mean and MDN = median represented in this dataset are based in the USA (N = 24) and contain a total of 1,008,220 items. Canada (M = 13,789) has three participating repositories containing a total of 41,366 items. The United Kingdom, Sweden and South Africa each have two participating repositories with a total of 61,289 items, 21,202 items and 8,783 items, respectively. South Africa has two participating repositories containing a total of 8,783 items. Australia and New Zealand each have one participating repository with 69,396 items and 12,207, respectively.

Of particular interest is the use ratio, which denotes how much IR content is accessed compared to how much is available. Use falls into three categories (Figure 2): IR with high numbers of items but low use ratio (11 repositories); IR with low numbers of items and low use ratio (17 repositories) and IR with low numbers of items but high use ratio (seven repositories). The first category (high item/low use ratio) has a total of 869,876 items from 11

Table 4. Number of repository items per country

Country	N	Min	MDN	M	Max	Total
The USA	24	2,095	29,046	42,009	124,436	1,008,220
Australia	1					69,396*
The United Kingdom	2	9,903	30,645	30,645	51,386	61,289
Canada	3	10,366	13, 021	13,789	17,979	41,366
Sweden	2	9,307	10,601	10,601	11,895	21,202
New Zealand	1					12,207*
South Africa	2	3,640	4,392	4,392	5,143	8,783

Note(s): N = institutional repository, M = mean and MDN = median, * indicates the subset of RAMP data used for the analysis. There is only one participating IR in those countries; therefore, we only present the total number of items for those countries

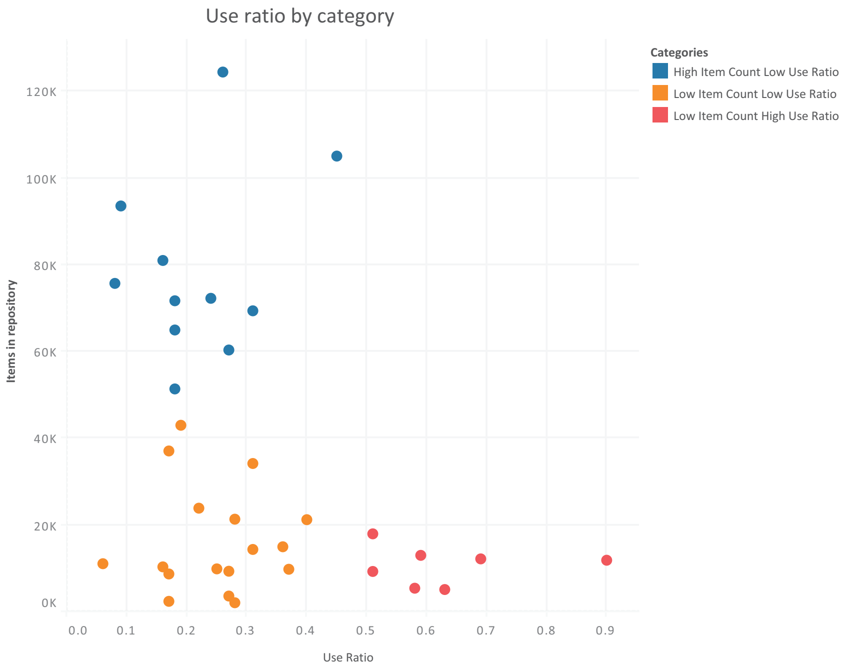

Figure 2. Use ratio by category

repositories (M = 79,080), and the mean use ratio is 0.22. The second category (low item/low use ratio) has a total of 277,589 items from 17 repositories (M = 16,328) and a mean use ratio of 0.25. The last category (low item/high use ratio) has a total of 74,998 items from seven repositories (M = 10,714) and a mean use ratio of 0.63. In the first two categories, during the five months from January to May 2019, on average, 25% or fewer of the items in the corresponding IR were accessed. On the other hand, in the last category, more than 60% of the items in the IR were accessed [3]. Complete statistics about the number of items per use ratio category and use ratio categories can be found in Tables 5 and 6.

In the first category, high item/low use (M = 79,080), the average number of items in each repository is 79,080 items. In the second category, low-item/low-use (M = 16,392), each repository has on average 16,392 items registered. The repositories in the low item/high use category (M = 10,714) have an average of 10,714 items. The total numbers of items in the first, second and third categories are 869,876, 277,589, and 74,998, respectively. The common pattern of use which can be seen from these statistics is that the average percentage of use is out of proportion with the total number items in the IR.

The further examination of the use ratio in each category supports the finding above that the use of the items in the IR is disproportionate to the total numbers of items within the IR. The first category has the most items on average and the lowest use ratio. In the first category, high-item/low-use (M = 0.22, SD = 0.1), during the five-month period, on average, only 22% (0.22) of items within these repositories were downloaded . Conversely, the low item/high use category has the fewest items on average but the highest use ratio. The average percentage of use per item for the last category, low item/high use (M = 0.63, SD = 0.14), is 63% (0.63).

The use ratio of each repository is visualized in Figure 3 and the use ratio by platform is shown in Table 7. The 35 repositories are reordered by use ratio in Table 8.

Use ratio appears to be positively affected by the number of ETDs in a given repository (Figure 4). The repositories that contain more ETD as a portion of their total items tend to have higher use ratios than repositories that contain fewer or no ETD.

Results of the country-device data analysis

In addition to the page-click data analyzed in the previous section, the RAMP also harvests daily search engine performance data describing the devices used to conduct searches on Google properties and the countries from which the searches originated. These data are aggregated in a combination of country and device, and they do not include URLs. Since it is

Category	N	Min	MDN	M	Max	Total
High items low downloads	11	51,386	72,275	79,080	124,436	869,876
Low items low downloads	17	2,095	11,095	16,329	43,008	277,589
Low items high downloads	7	5,143	11,895	10,714	17,979	74,998

Note(s): N = repository, M = mean and MDN = median

Table 5. Number of items by use ratio category

Category	N	Min	MDN	M	SD	Max
High items low downloads	11	0.08	0.18	0.22	0.11	0.45
Low items low downloads	17	0.06	0.27	0.25	0.09	0.40
Low items high downloads	7	0.51	0.59	0.63	0.14	0.90

Note(s): N = repository, M = mean, MDN = median and SD = standard deviation

Table 6. Use ratio category

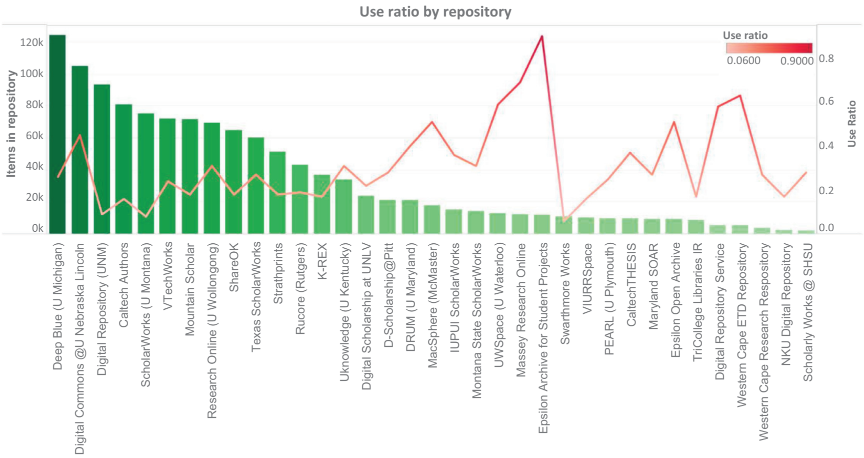

Figure 3.
Use ratio of each repository

Platform	*N*	Min	MDN	*M*	SD	Max
EPrints	6	0.16	0.32	0.40	0.28	0.90
Fedora	2	0.19	0.38	0.38	0.28	0.58
DSpace	20	0.16	0.27	0.32	0.16	0.69
Digital Commons	7	0.06	0.22	0.22	0.15	0.45

Note(s): *N* = repository, *M* = mean, MDN = median and SD = standard deviation

Table 7.
Use ratio by platform

Repository name	#Items in repository	#Unique CCD items	Use ratio
Epsilon Archive for Student Projects	11,895	10,753	0.90
Massey Research Online	12,207	8,445	0.69
Western Cape ETD Repository	5,143	3,244	0.63
UWSpace (*U* Waterloo)	13,021	7,657	0.59
Digital Repository Service (Northeastern)	5,446	3,173	0.58
MacSphere (McMaster)	17,979	9,122	0.51
Epsilon Open Archive	9,307	4,747	0.51
Digital Commons @*U* Nebraska Lincoln	105,065	47,418	0.45
DRUM (*U* Maryland)	21,246	8,437	0.40
CaltechTHESIS	9,814	3,673	0.37
IUPUI ScholarWorks	15,000	5,379	0.36
Montana State ScholarWorks	14,381	4,405	0.31
UKnowledge (*U* Kentucky)	34,196	10,628	0.31
Research Online (*U* Wollongong)	69,396	21,343	0.31
Scholarly Works @ SHSU	2,095	586	0.28
D-Scholarship@Pitt	21,358	6,011	0.28
Texas ScholarWorks	60,359	16,255	0.27
Western Cape Research Repository	3,640	972	0.27
Maryland SOAR	9,359	2,543	0.27
Deep Blue (*U* Michigan)	124,436	31,773	0.26

(continued)

Table 8.
Repositories sorted by use ratio

Repository name	#Items in repository	#Unique CCD items	Use ratio
PEARL (*U* Plymouth)	9,903	2,449	0.25
VTechWorks	72,275	17,610	0.24
Digital Scholarship at UNLV	23,896	5,339	0.22
RUcore (Rutgers)	43,008	8,264	0.19
Mountain Scholar	71,708	12,611	0.18
ShareOK	64,972	11,829	0.18
Strathprints	51,386	9,451	0.18
TriCollege Libraries IR	8,728	1,474	0.17
K-REX	37,084	6,433	0.17
NKU Digital Repository	2,420	408	0.17
Caltech Authors	81,000	13,183	0.16
VIURRSpace	10,366	1,670	0.16
Digital Repository (UNM)	93,564	8,839	0.09
ScholarWorks (*U* Montana)	75,715	5,836	0.08
Swarthmore Works	11,095	687	0.06

Table 8.

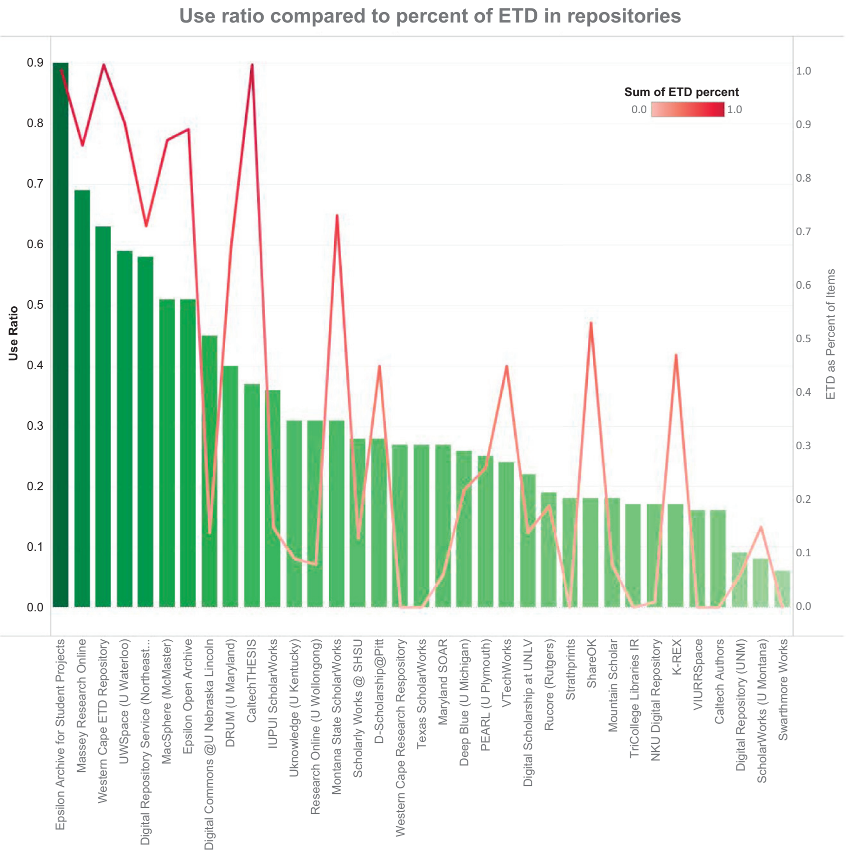

Figure 4.
Use ratio compared to percent of ETD in repositories

not possible to filter clicks on citable content URLs from clicks on HTML URLs, the following analysis describes trends in search engine performance per IR rather than per item.

RQ2. How do devices used to access IR content affect user behavior?

Table 9 shows that most of the clicks on IR content (70.3%) came from desktop users. Tablets were used much less frequently than desktop and mobile devices.

Figure 5 and Table 10 show how device use varies depending on the SERP position of IR pages. The number of clicks originating from desktops, mobile devices and tablets in the five-month period was 7,438,457, 2,878,834 and 263,205, respectively. The “position” relates to position in the SERP where each page contains ten results; a lower position indicates a better placement in the SERP. When desktop interfaces were used to conduct Google searches, each item in the IR was, on average, downloaded 3.13 times and the median position of the IR content in the SERP was 60.2. The average download of an item when mobile phones were used to search for information in the repositories was 2.04, and the median position of the IR content in the SERP was 18.7. The average number of downloads of an item when tablets were used to search for information in the repositories was 0.48, and the median position of the IR content in the SERP was 11.9.

Analysis of access to IR content by device, country and Global North/South origin was performed by merging RAMP country-device data with a manually tabulated dataset of country names, three letter ISO 3166 codes (International Organization for Standardization, 2020; Wikipedia Contributors, 2020) and Global North/South designations per country (Meta Contributors, 2020). The RAMP dataset contains search engine result data from “unknown

Device	Count of occurrences	Total clicks	% of total clicks
Desktop	2,372,958	7,438,457	70.30
Mobile	1,410,171	2,878,834	27.21
Tablet	550,301	263,205	2.49

Table 9. Breakdown of clicks by device

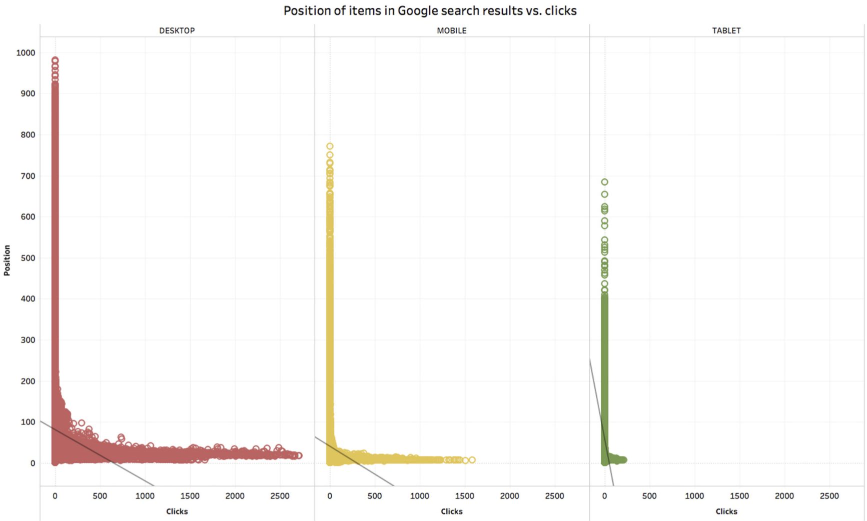

Figure 5. Position of items in Google search results vs clicks

regions," and since there is no way to tell whether users conducting the corresponding searches came from either the Global North or South, these data were dropped from the following analyses. The total number of dropped rows and click sums is documented in (Wheeler *et al.*, 2020a).

As indicated by Figure 6, users in the Global North generally accessed IR more frequently than those in the Global South. Within the five-month period, 57% of clicks (5,993,841) in the data came from users in the Global North and 43% of clicks (4,562,586) came from users in the Global South (Table 11). In total, three of the top five countries with the most clicks are in the Global North: USA (2,763,548); United Kingdom (590,516) and Canada (421,358). The

Table 10.
Positions vs clicks

Variable	*N*	Min	MDN	*M*	SD	Max
Desktop	2,372,958					
Clicks		0	0	3.13	34.59	3,258
Position		1	60.2	89.4	74	982
Mobile	1,410,171					
Clicks		0	0	2.04	28.51	22,740
Position		1	18.7	45.3	58.2	771
Tablet	550,301					
Clicks		0	0	0.48	3.95	211
Position		1	11.9	45.26	67.2	685

Note(s): N = device use frequency, M = mean (page position), SD = standard deviation and MDN = median

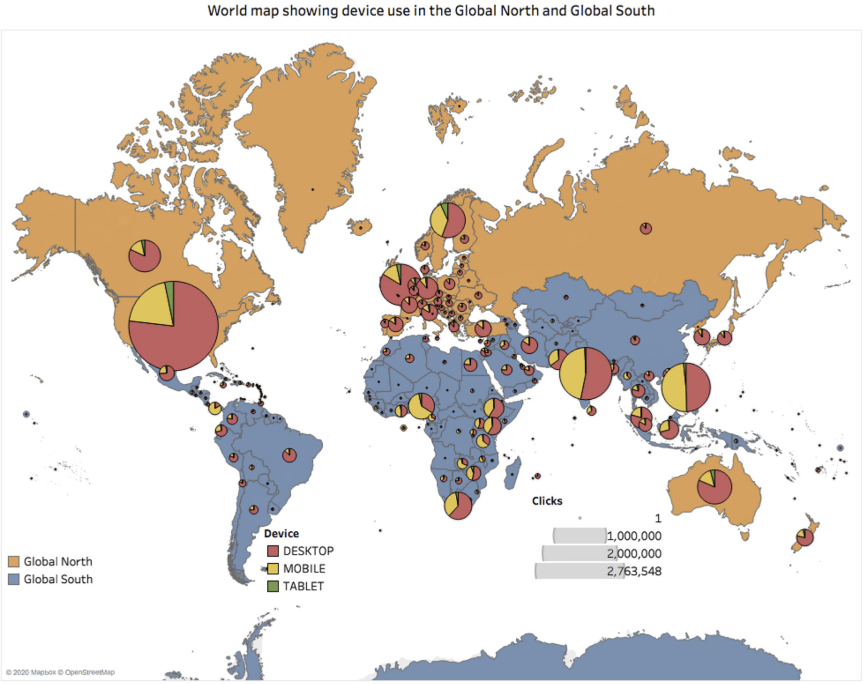
World map showing device use in the Global North and Global South
Global North
Global South
Device
DESKTOP
MOBILE
TABLET
Clicks
1
1,000,000
2,000,000
2,763,548
© 2020 Mapbox © OpenStreetMap

Figure 6.
World map showing device use in the Global North and Global South

Global South has two countries in the top five: India (950,106) and the Philippines (820,415) (Table 12).

Table 9 showed that desktop operating systems were consistently the dominant devices used to download and read documents (70.3%), followed by mobile phones (27.2%). Tablets accounted for only a small portion of use (2.5%). Table 13 shows that users in the Global North accounted for 79% of the clicks from desktops, 18% from mobile phones and 3% from tablets, while users in the Global South accounted for 59% of the clicks from desktops, 39.4% of the clicks from mobile phones and 1.6% from tablets.

Discussion

Limitations

A limitation of RAMP data is the reliance of the service on GSC as the sole data source. While the current literature demonstrates that the vast majority of IR traffic comes from Google properties (Macgregor, 2019), content is sometimes shared through social media applications like Twitter and Facebook as well as academic professional networks like ResearchGate and Academia.edu and networked research services like Mendeley. In addition, links to IR content may be embedded in public relations pages, blogs and news stories, course management software, etc. RAMP can only capture clicks on IR content links embedded in these services and pages if they are indexed and exposed via Google SERP. RAMP data may therefore be considered to be somewhat conservative, but the authors contend that a full accounting of all clicks on IR content is far less important than a substantial dataset with assurance that nearly all clicks are human generated. As discussed in the literature review section, Google's highly successful pay-per-click advertising model (Alphabet, Inc., 2015) depends on its ability to filter out robot traffic, so there is high confidence that the data provided through RAMP represent human traffic.

Location	Clicks	Percent
Global North	5,993,841	57%
Global South	4,562,586	43%

Table 11. Clicks from users in the Global North and the Global South

#	Country	Location	Clicks
1	The USA	Global North	2,763,548
2	India	Global South	950,106
3	The Philippines	Global South	820,415
4	The United Kingdom	Global North	590,516
5	Canada	Global North	421,358

Table 12. Top five countries that generated IR traffic to RAMP-registered repositories

Location	Desktop	Mobile	Tablet
Global North	4,731,597 (78.9%)	1,070,754 (17.9%)	191,490 (3.2%)
Global South	2,692,760 (59%)	1,798,312 (39.4%)	71,514 (1.6%)

Table 13. Device use between users in the Global North and the Global South

A limitation of the page-click analysis is the difference between collection dates of information including IR item counts and the separate count of ETD that occurred for each IR. ETD counts were collected during the first ten days of June 2019 while the RAMP dataset reaches only to the end of May 2019, so it is likely that some ETD were added to one or more RAMP IR in the interim. The difference is very small and does not significantly affect the results of the analyses.

A second limitation of the page-click and use ratio analysis is that the scale of the dataset required the development of an automated method for deriving the ratio's numerator. For each repository, this number is the count of unique items that contain content files with positive click values in RAMP. "Items" is here understood as the HTML pages that include the metadata and content for individual IR objects and are therefore the parent pages of the citable content URLs tracked by RAMP. HTML URLs of the parent pages have to be inferred or reconstructed using information contained within the content file URLs. The process for doing so is platform specific, and in the case of EPrints and Fedora it is also limited for identifying the parent pages of PDF files rather than all content types. Variation among platforms may result in a case where the use ratio is not an equivalent method of comparison between IR that use different platforms. Even so, the authors are confident that the method is accurate based on data integrity checks built into the process as described in available documentation (Wheeler *et al.*, 2020a).

Size of IR

There is a large variance in the number of items in the 35 IR that comprised our dataset. The smallest repository contains just over 2,000 items and the largest nearly 125,000; the median was almost 18,000 items. The size of the IR did not always align with the size of the institution. For example, Caltech Authors is the fourth largest IR in the dataset with 81,000 items, but Caltech itself has only 300 professorial faculty, 600 research scholars and approximately 2,250 undergraduate and graduate students (California Institute of Technology, 2019). Caltech is therefore a very high performing institution when measured by research publications and one where most of the publications appear to be deposited in the IR.

Conversely, some larger institutions fall near the lower end of the range in terms of number of items in their repositories. The fact that some small institutions have large repositories while some large institutions have small repositories may be attributed to several factors including the research and publishing culture on campus, the success of the library in attracting participation in the IR and even the platform itself. For example, the Digital Commons platform from BePress can also be used as a journal publishing platform, which can result in large numbers of articles published in locally managed journals.

Use ratio

The use and performance of the IR also vary significantly and can be categorized into three groups. We calculated use ratio as the number of unique items containing content file URLs with positive click values in RAMP data divided by the total number of items in the repository. Using this calculation, we have shown that some IR have high numbers of items but comparatively low use ratio (<0.3); others have low numbers of items and low use ratio (<0.4); and yet other IR have low numbers of items, but the use ratio is relatively high (>0.7) compared to other institutions in the set. This shows that larger IR do not necessarily experience higher rates of use than smaller IR and that other factors may be at play. Factors affecting the ratio of use could include whether the repository has been successfully harvested and indexed by Google and Google Scholar, position of the items in the SERP, attractiveness of the item title and rich snippet that appear in the SERP, the intellectual content of the items and even current trends in research.

In general, it must be noted that nearly all of the IR in this dataset suffer from low use as indicated by downloads, although, admittedly, our use ratio is a rather blunt instrument of measure. For example, a gross simplification would be to say that 31% of the 14,381 items in the Montana State University ScholarWorks Repository were accessed at least once. While this is true as a general assessment of the repository's usage, it is likely that a more granular analysis would demonstrate that the majority of access and use of IR content is driven by a small number of very popular items that are downloaded many times. Conversely, there are probably many items that are seldom accessed or never downloaded at all. Quantifying detailed usage is possible by examining individual URIs in the RAMP dataset. The RAMP team has completed a preliminary analysis of the data along these lines, but the discussion of these initial findings is beyond the scope of this paper [4]. For now, satisfaction must lie in the realization that the same use ratio calculation was applied to every repository.

The different platforms also showed some difference in use ratio. Digital Commons repositories had the lowest overall use ratio of the four platforms in this dataset. Whether this is due to the capacity of Digital Commons to facilitate the archiving of materials that are not research publications or ETD is a question that cannot be answered with our current dataset.

The effect of ETD on use ratio

The Epsilon Archive for Student Projects, which primarily houses ETD, showed the highest use ratio in our dataset. Indeed, when one draws back the lens to determine whether there is any common characteristic that positively affects use ratios, the concentration of ETD in repositories emerges as a considerable factor. In addition to the Epsilon Archive for Student Projects, Massey Research Online, Western Cape ETD Repository, UWSpace and Caltech THESES are among the repositories whose content consists almost entirely of ETD and whose use ratios are the highest in our dataset, ranging from 0.37 to 0.90. For comparison, we can again use the example of Caltech Authors, where, despite a large repository of 81,000 items, it showed one of the lowest use ratios of only 0.16.

Why do repositories with high concentrations of ETD seem to experience more use than repositories that consist mainly of faculty publications or other kinds of items? Is it because theses and dissertations are less likely to be published anywhere else, except perhaps in the fee-based ProQuest® *Dissertations and Theses Global* database? Is it that theses and dissertations represent new and original research that might be of interest to researchers, corporations or even governments? Here again lies another avenue for future research that can be facilitated by RAMP's open dataset.

Discoverability

Multiple factors can affect the discoverability and use of IR items. RAMP data are confirming years of speculation that many IR suffer from low use and one significant causal factor is likely to be low indexing ratios in search engines. Search engine optimization (SEO) tends to be inconsistently practiced or even nonexistent in libraries, and this can make it difficult for search engines to uniformly harvest and index IR. If items in the repository do not appear in the various Google search indexes, then there is no possibility of them being surfaced in the Google SERP and no possibility of RAMP data showing downloads.

As mentioned in the discussion on use ratio, another potentially significant factor in discoverability is the position where the item appears in the SERP. Items that appear within the first few pages of SERP logically have a much higher chance of being downloaded from a repository than items that appear further down the list. Position may be affected by SEO practices, including metadata that can help the search engine determine how relevant the item is to the user's query.

Device use

RAMP data highlight how devices may affect the behavior of users who search and access IR content. Despite the ubiquity of mobile devices, our dataset shows that most users still employ desktop operating systems (including laptop computers) and that they tend to delve further into the SERP and download more items from IR than mobile or tablet users. It is possible that the user experience with mobile devices may be degraded enough in both search engine and IR interfaces to result in these lower numbers. Or perhaps researching and writing is simply more integrated and convenient on a desktop. For example, a researcher working in a desktop environment may search for articles in a browser, while simultaneously using a reference manager and a word processing application in other windows. These applications may even work in concert through plug-ins, which provide a level of convenience and sophistication that is simply not available on mobile or tablet devices.

Theories aside, the fact remains that many people have only mobile devices to satisfy their research needs. In fact, nearly five billion people worldwide were estimated to own mobile devices in 2019 and "over half of these connections are smartphones" (Taylor and Silver, 2019). By one measure, worldwide distribution of mobile devices is approximately 52%, while desktops are 45% and tablets are 3%, with the tilt toward mobile devices being more pronounced in the Global South (StatCounter GlobalStats, 2019). The data in this study show that mobile device users in the Global South access the IR content at a much higher rate (39.4%) than mobile device users in the Global North (17.9%). When paired with the search behavior data of desktop users, it is reasonable to infer that users in the Global South are at a disadvantage because mobile devices are more limiting for in-depth research than desktops.

Conclusion

We wish to emphasize that the conclusions drawn from this study should be considered with caution. This work still represents only 35 repositories from a worldwide landscape that may exceed 4,500 (University of Southampton, 2019). More participation in the RAMP will help to grow the dataset available for analysis.

The RAMP dataset holds much potential for further research. Some examples include the following:

(1) Analyze the scholarly record in the IR to better understand what is available and what users seek.

(2) Identify and correct SEO or other problems for repositories that experience low use.

(3) Support or dispute the theory that developing countries benefit from open access to research publications in the IR.

(4) Analyze metadata from articles with high SERP positions to reveal characteristics that foster discoverability and use.

(5) Test the theory that IR is actually supplying preprint content for citations of articles behind paywalls.

The purpose of this research was to demonstrate the kind of analysis that is possible with the openly published RAMP dataset and to encourage its use for further research. Future research holds promise of a more nuanced picture of information-seeking behaviors of users.

Notes

1. RAMP website – https://rampanalytics.org

2. The higher average number of items may be due to the fact that many customers take advantage of Digital Commons' journal publishing features.

3. Use ratio as defined here is a general indicator of how much of IR content is accessed via SERP. Additional descriptive statistics available from (Wheeler *et al.*, 2020a, b) demonstrate that in most cases overall use is driven by a smaller percentage of highly accessed items.

4. The RAMP_summary_stats__20200907.csv file available from this study's GitHub repository (Wheeler *et al.*, 2020a) contains summary statistics demonstrating that a majority of IR use is driven by a few highly accessed items.

   In the interest of transparency, data sharing and reproducibility, the author(s) of this article have made the data underlying their research openly available. It can be accessed by following the link here: https://rampanalytics.org

References

Almind, T.C. and Ingwersen, P. (1997), "Informetric analyses on the world wide web: methodological approaches to webometrics", *Journal of Documentation*, Vol. 53 No. 4, pp. 404-426.

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

Arlitsch, K. and Grant, C. (2018), "Why so many repositories? Examining the limitations and possibilities of the institutional repositories landscape", *Journal of Library Administration*, Vol. 58 No. 3, pp. 264-281.

Arlitsch, K., Askey, D. and Wheeler, J. (2019), "Analyzing aggregate IR use data from RAMP", *PowerPointpresented at the Open Repositories 2019*, Hamburg, Germany, 11 June, doi: 10.5281/zenodo.3243348 (accessed 9 February 2020)..

Björneborn, L. and Ingwersen, P. (2004), "Toward a basic framework for webometrics", *Journal of the American Society for Information Science and Technology*, Vol. 55 No. 14, pp. 1216-1227.

Bruns, T. and Inefuku, H.W. (2015), "Purposeful metrics: matching institutional repository metrics to purpose and audience", in Callicott, B.B., Scherer, D. and Wesolek, A. (Eds), *Making Institutional Repositories Work*, Purdue University Press, West Lafayette, pp. 213-234.

California Institute of Technology (2019), *Caltech at a Glance*, Caltech, Educational, 8 August, available at: https://www.caltech.edu/about/at-a-glance (accessed 6 October 2019).

Fralinger, L. and Bull, J. (2013), "Measuring the international usage of US institutional repositories", *OCLC Systems and Services: International Digital Library Perspectives*, Vol. 29 No. 3, pp. 134-150.

Frost, A. (2019), "The ultimate guide to Google search console in 2019", *HubSpot*, 5 September, available at: https://blog.hubspot.com/marketing/google-search-console (accessed 16 November 2019).

Google, Inc. (2020), "Search console APIs", *Google Developers*, available at: https://developers.google.com/webmaster-tools/search-console-api-original/ (accessed 7 January 2020).

Greene, J. (2016), "Web robot detection in scholarly open access institutional repositories", *Library Hi Tech*, Vol. 34 No. 3, pp. 500-520.

Harnad, S. and McGovern, N. (2009), "Topic 4: institutional repository success is dependent upon mandates", *Bulletin of the American Society for Information Science and Technology*, Vol. 35 No. 4, pp. 27-31.

International Organization for Standardization (2020), *ISO 3166 Country Codes*, ISO, Non-Governmental Organization, available at: https://www.iso.org/iso-3166-country-codes.html (accessed 12 April 2020).

IRUS-UK Team (2013), *IRUS-UK Position Statement on the Treatment of Robots and Unusual Usage*, November, available at: https://irus.jisc.ac.uk/documents/IRUS-UK_position_statement_robots_and_unusual_usage_v1_0_Nov_2013.pdf (accessed 25 August 2019).

Jisc (2019), *Welcome to IRUS-UK*, IRUS-UK, Government, available at: https://irus.jisc.ac.uk (accessed 11 November 2019).

Kim, K. (2018), "DLF-Jisc Pilot Project webinar", 23 March, available at: https://www.diglib.org/recording-available-for-irus-usa-webinar/ (accessed 25 August 2019).

Lagzian, F., Abrizah, A. and Wee, M.-C. (2015), "Measuring the gap between perceived importance and actual performance of institutional repositories", *Library and Information Science Research*, Vol. 37 No. 2, pp. 147-155.

Macgregor, G. (2019), "Improving the discoverability and web impact of open repositories: techniques and evaluation", *Code4Lib Journal*, No. 43, available at: https://journal.code4lib.org/articles/14180 (accessed 25 August 2019).

MacIntyre, R. and Jones, H. (2016), "IRUS-UK: improving understanding of the value and impact of institutional repositories", *The Serials Librarian*, Vol. 70 Nos 1-4, pp. 100-105.

McDonald, R.H. and Thomas, C. (2008), "The case for standardized reporting and assessment Requirements for institutional repositories", *Journal of Electronic Resources Librarianship*, Vol. 20 No. 2, pp. 101-109.

Meta Contributors (2020), *List of Countries by Regional Classification*, Wikimedia Meta-Wiki, Meta, discussion about Wikimedia projects, 1 April, available at: https://meta.wikimedia.org/w/index.php?title=List_of_countries_by_regional_classification&oldid=19943813 (accessed 12 April 2020).

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.

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

OBrien, P., Arlitsch, K., Mixter, J., Wheeler, J. and Sterman, L.B. (2017), "RAMP – the Repository Analytics and Metrics Portal: a prototype web service that accurately counts item downloads from institutional repositories", *Library Hi Tech*, Vol. 35 No. 1, pp. 144-158.

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*, Vol. 12 No. 11, doi: 10.1045/november2006-organ.

Parulian, N. (2020), *RAMP Data Viz*, Tableau Public, Tableau Public, 9 February, available at: https://tinyurl.com/rsxjdv8 (accessed 9 February 2020).

Perrin, J.M., Yang, L., Barba, S. and Winkler, H. (2017), "All that glitters isn't gold: the complexities of use statistics as an assessment tool for digital libraries", *The Electronic Library*, Vol. 35 No. 1, pp. 185-197.

Rodrigues, E., Bollini, A., Cabezas, A., Castelli, D., Carr, L., Chan, L., Humphrey, C., Johnson, R., Knoth, P., Manghi, P., Matizirofa, L., Perakakis, P., Schirrwagen, J., Selematsela, D., Shearer, K., Walk, P., Wilcox, D. and Yamaji, K. (2017), "Next generation repositories: behaviours and technical recommendations of the coar next generation repositories working group", *Zenodo*. doi: 10.5281/ZENODO.1215014.

StatCounter GlobalStats (2019), *Desktop vs Mobile vs Tablet Market Share Worldwide, Oct 2018–Oct 2019*, StatCounter GlobalStats, October, available at: https://gs.statcounter.com/platform-market-share/desktop-mobile-tablet/worldwide (accessed 29 November 2019).

Taylor, K. and Silver, L. (2019), *Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally*, Pew Research Center, Washington, District Columbia, available at: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ (accessed 29 November 2019).

Thompson, S., Lambert, J., Macintyre, R., Chaplin, D., Jones, H., Wong, L., Perrin, J., Rubinow, S., Kim, K., Nowviskie, B., Needham, P., Williford, C. and Graham, W. (2019), "Bringing IRUS to the USA: international collaborations to standardize and assess repository usage statistics", *Proceedings of the 2018 Library Assessment Conference: Building Effective, Sustainable, Practical Assessment*, 5–7 December 2018, Houston, TX, presented at the Library Assessment Conference – Building Effective, Sustainable, Practical Assessment, Association of Research Libraries, pp. 564-577.

University of Southampton (2019), *Registry of Open Access Repositories*, Educational, available at: http://roar.eprints.org (accessed 6 October 2019).

Wheeler, J., Arlitsch, K., Parulian, N. and Pham, M. (2020a), "RAMP analyses scripts, R", available at: https://github.com/imls-measuring-up/ramp-analyses-scripts (accessed 9 May 2020).

Wheeler, J., Arlitsch, K., Pham, M. and Parulian, N. (2020b), *RAMP Data Subset, January 1 through May 31, 2019*, University of New Mexico, 14 January, doi: 10.5061/dryad.fbg79cnr0.

Wikipedia Contributors (2020), *ISO 3166-1 Alpha-3*, Wikipedia: The Free Encyclopedia, 14 March, available at: https://en.wikipedia.org/w/index.php?title=ISO_3166-1_alpha-3&oldid=945527106 (accessed 12 April 2020).

About the authors

Kenning Arlitsch has been dean of the library at Montana State University since 2012. He has held positions as an instruction librarian, in digital library development, IT services and administration. His funded research has focused on SEOas well as measuring impact and use of digital repositories. Kenning holds a MLIS from the University of Wisconsin-Milwaukee and a Ph.D. in library and information science from Humboldt University in Berlin, Germany. His dissertation on Semantic Web Identity examined how well research libraries and other academic organizations are understood by search engines. Kenning Arlitsch is the corresponding author and can be contacted at: kenning.arlitsch@montana.edu

Jonathan Wheeler is a Data Curation Librarian within the University of New Mexico's College of University Libraries and Learning Sciences. Jon's role in the Libraries' Data Services initiatives includes the development of research data ingest, packaging and archiving workflows. His research interests include workflow development in support of quality control and streamlined data storage, dissemination, archiving and preservation. Jon holds an M.S. in library science from the University of Illinois at Urbana-Champaign.

Minh Thi Ngoc Pham is currently a Ph.D. candidate at the School of Information Science and Learning Technologies at the University of Missouri, Columbia. She holds a master's degree in Globalization and Education Change from Lehigh University, Pennsylvania. Minh's research interests include game-based learning with Virtual Reality (VR) and Augmented Reality (AR) tools and geographic information system (GIS) for research and decision-making. She was a fellow in Drexel University's LIS Education and Data Science (LEADS) program in 2019.

Nikolaus Nova Parulian is a Ph.D. student in the School of Information Sciences at the University of Illinois at Urbana-Champaign. He also holds a Master of Science in Information Management (MSIM) from the University of Illinois. His research interests include topics related to machine learning, text mining, data quality management, and network analysis. Nikolaus was a fellow in Drexel University's LIS Education and Data Science (LEADS) program in 2019.