2020

Repository Analytics and Metrics Portal (RAMP) Workflow Documentation and Data Definition

Jon Wheeler
University of New Mexico - Main Campus, jwheel01@unm.edu

Kenning Arlitsch
Montana State University

Repository Analytics and Metrics Portal (RAMP) Workflow Documentation and Data Definition

JONATHAN WHEELER, DATA CURATION LIBRARIAN, UNIVERSITY OF NEW MEXICO
https://orcid.org/0000-0002-7166-3587

KENNING ARLITSCH, DEAN OF THE LIBRARY, MONTANA STATE UNIVERSITY
https://orcid.org/0000-0002-5919-735X

Abstract

The Repository Analytics & Metrics Portal (RAMP) is a web service that leverages Google Search Console (GSC) data to provide a set of baseline search engine performance metrics for a global, cross-platform group of institutional repositories (IR). Since launching in 2017, RAMP has grown from 3 to more than 50 participating repositories. The underlying data are unique in scope and size, and offer many opportunities for novel analyses of IR search engine performance. The data may be augmented to enable additional analyses including metadata mining and bibliometrics. In November 2019, the RAMP team released a publicly available subset of the RAMP dataset, consisting of daily GSC data for 35 participating repositories harvested between January 1 and May 31, 2019. The purpose of this article is to provide information and increased transparency about how RAMP data are harvested, processed, and audited for quality control. This article is also intended to serve as more extensive, complementary documentation for the published dataset and any published research findings that use RAMP data.

Introduction

The Repository Analytics and Metrics Portal (RAMP) was launched in January 2017 as a prototype implementation of a new model for reporting institutional repository (IR) metrics (OBrien et al, 2016; OBrien et al., 2017). The model aggregates data from Google Search Console (GSC) in a novel way to provide information about IR search performance and user activity that investigators believe provides detailed IR access and use information that is not currently available through other means.

In addition to offering an alternative model of IR performance metrics, RAMP as a platform has demonstrated further benefits including:

- A unique aggregation of GSC data for various IR, across which data collection and analysis methods are consistently applied regardless of IR platform, size, etc. This aggregation enables comparative analysis of IR performance across institutions that is not otherwise possible among participants.

- Data persistence. Whereas GSC formerly maintained data for 90 days at most, to date RAMP has been able to maintain all data for all participating IR. For early participants, that means RAMP data may go back as far as three years. This enables a long term analysis that is also unique to RAMP participants.
- Simplicity. The process for joining RAMP is fast and requires no platform configuration changes.

Although RAMP was initially conceived as a resource for individual IR managers, the combination of cross-platform data aggregation and data persistence have resulted in a dataset that is unique in size and scope. As of January, 2020, the RAMP dataset consists of over 618 million "rows"[1] representing search engine performance data for 54 production IR from 5 continents. Platforms in use by RAMP repositories include DSpace, Digital Commons, EPrints, Fedora, as well as custom built solutions for cultural heritage and data repositories. The size and type of participating institutions varies from multi-institution consortia, state, and R1 universities to technical institutes and smaller, private institutions. To the best of our knowledge, no other datasets comparable to RAMP exist which offer similar potential to the global IR community for long term and comparative analysis of repository search engine performance, access, and use. Other potential analytic applications of the data include the development of methods for automated IR search engine optimization, metadata standardization and augmentation, and the development of machine learning algorithms.

The RAMP team anticipates that the community will recognize additional research avenues, and has released a subset consisting of complete RAMP data for 35 participating repositories for the five month period January 1 through May 31, 2019 (Wheeler et al., 2019). This paper is intended to serve as additional, detailed documentation of RAMP and the RAMP dataset to provide transparency about the RAMP service and to support use and analysis of the data.

Limitations

It is a goal of the RAMP team to publish the complete dataset. To date we have opted to publish a subset due to limitations including

- Temporary configuration errors among participating IR that may result in incomplete data for those IR for a given time period.
- Limited consent from participating IR. As will be seen in the documentation provided below, RAMP data do not contain any sensitive or personally identifiable data. However, there may be sensitivities around institutional or library reputation depending on how data are analyzed or compared. The RAMP team greatly values and respects the interests of participating IR and actively seeks permission from participants to include their data in any public release.

[1] RAMP data are not stored in tabular form, but are instead indexed in a hosted instance of Elasticsearch. Search engine result page data retrieved from GSC are indexed as 'documents' in Elasticsearch. In the case of RAMP data, each 'document' is analogous to a row of tabular data so the more familiar term is used throughout this article.

It should be noted that while both limitations affect the number of repositories whose data are included in the released subset, these limitations do not impact the completeness of the data published for any specific IR. For example, the University of New Mexico's Digital Repository is one of the 35 IR whose data are included in the published subset, and the subset includes 100% of UNM's RAMP data from January 1 through May 31, 2019. Alternatively, if a RAMP participating IR experienced a configuration problem during that time, then the subset does not include any data at all for that IR regardless of whether the error has since been resolved.

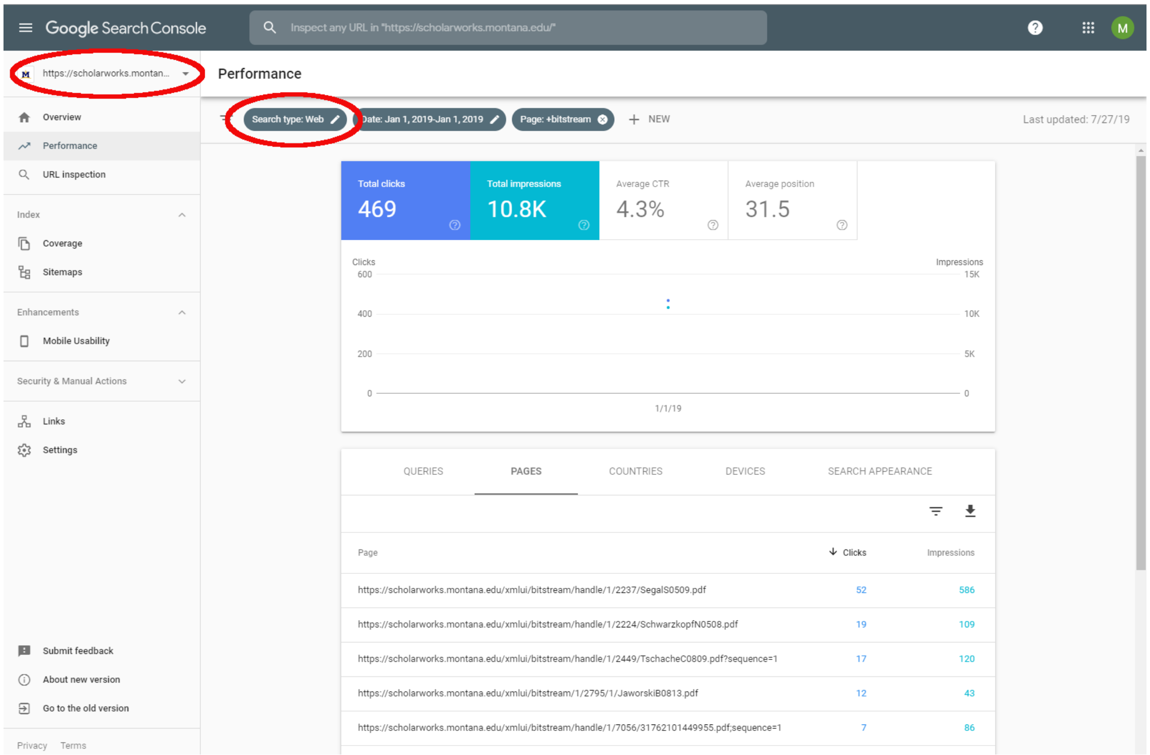
The RAMP Workflow: A Manual Demonstration Using Google Search Console

RAMP data are retrieved from the Google Search Console API. As described in OBrien (2016) and OBrien (2017), GSC data differ from other metrics including those reported by Google Analytics in that GSC data can include significantly more information about the appearance of IR content in search engine result pages (SERP) for searches performed on Google properties including web search and Google Scholar.

Because RAMP uses data from GSC, metrics and results similar if not identical to those reported by RAMP can be retrieved by participating IR managers using their repository's GSC dashboard. The value addition that RAMP provides is the concatentation, aggregation and processing of data from mutiple GSC API endpoints, each of which must currently be accessed as separate reports when using the GSC web interface. For example, IR search engine performance data as provided in the GSC dashboard are separated by the root repository URL and search type. Repositories may have multiple root URLs, as in the case of IR that expose content via both HTTP and HTTPS protocols. Although many HTTP pages will redirect to HTTPS and recent changes to GSC provide further means of consolidating data, managers of IR that use both HTTP and HTTPS URLs must generally view data for each protocol separately when looking at their GSC dashboards.

Within GSC, SERP data for searches performed via the standard Google web search interface (https://www.google.com/) are also reported separately from data for searches performed via the "images" search interface (https://images.google.com/). Since IR search engine performance data for web and image search results must be separately requested for both HTTP and HTTPS instances of an IR, within the online GSC dashboard this can amount to navigating to four different reports to retrieve complete data for any given time period.

Finally, when measuring access to IR content it is useful to distinguish clicks on actual content files (PDF, CSV, TIF, etc.) from clicks on HTML files. Since GSC dashboard reports provide metrics about clicks on all page types by default, limiting the data to clicks on content files requires that a *page* filter be applied to filter results to include only clicks on content files. In the case of a DSpace repository, the filter term for content file URLs is "bitstream." Figures 1 through 4 demonstrate the individual steps that must be taken and the four separate reports that would have to be viewed using Montana State University's IR GSC dashboard in order to get a sum of click counts on content file URLs for a single day, January 1, 2019. For easier comparison with the RAMP dashboard, Table 1 tabulates the "Total clicks" on content files from each report. RAMP consolidates these steps to produce a single report, as shown in Figure 5.

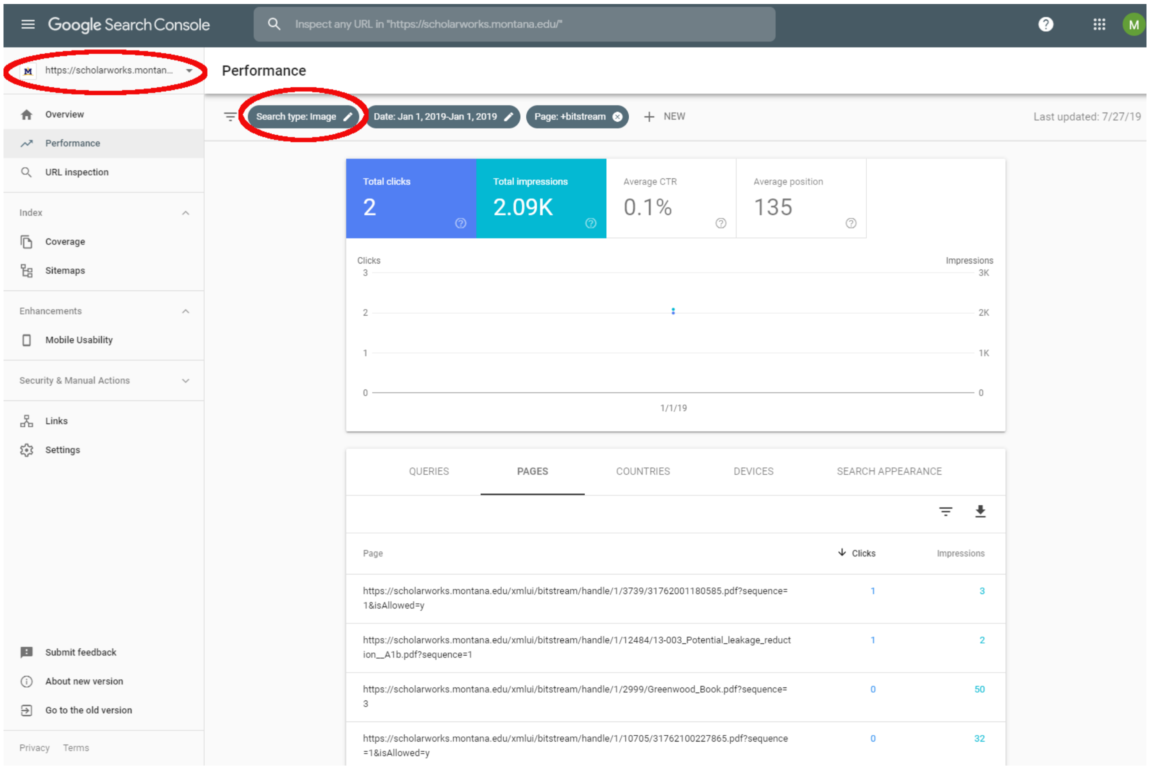


Search type: Web
Date: Jan 1, 2019-Jan 1, 2019
Page: +bitstream
Last updated: 7/27/19

Total clicks 469
Total impressions 10.8K
Average CTR 4.3%
Average position 31.5

Page	Clicks	Impressions
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/2237/SegalS0509.pdf	52	586
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/2224/SchwarzkopfN0508.pdf	19	109
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/2449/TschacheC0809.pdf?sequence=1	17	120
https://scholarworks.montana.edu/xmlui/bitstream/1/2795/1/JaworskiB0813.pdf	12	43
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/7056/31762101449955.pdf;sequence=1	7	86

Figure 1: MSU ScholarWorks IR GSC Dashboard for January 1, 2019. A "page" filter has been applied to include only URLs with 'bitstream' in the path. Note the HTTPS root URL and the search type, circled.



https://scholarworks.montan...

Performance

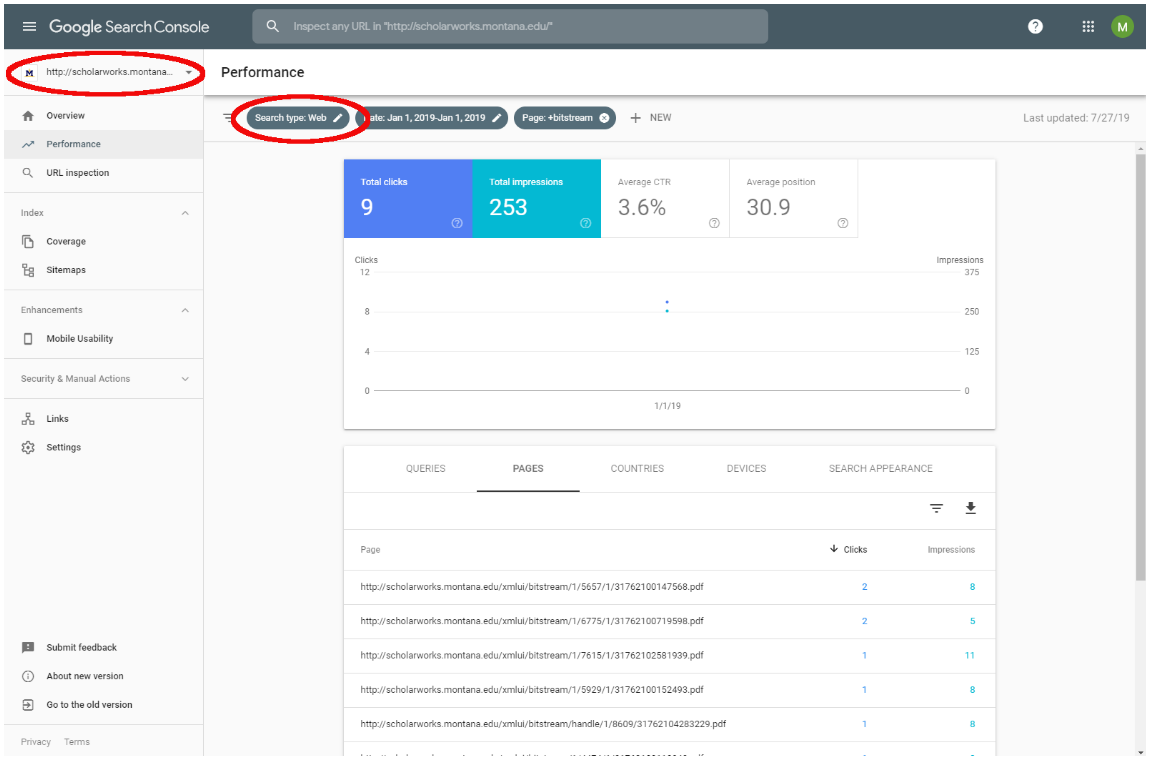
Search type: Image | Date: Jan 1, 2019-Jan 1, 2019 | Page: +bitstream

Last updated: 7/27/19

Total clicks 2 | Total impressions 2.09K | Average CTR 0.1% | Average position 135

Page	Clicks	Impressions
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/3739/31762001180585.pdf?sequence=1&isAllowed=y	1	3
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/12484/13-003_Potential_leakage_reduction__A1b.pdf?sequence=1	1	2
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/2999/Greenwood_Book.pdf?sequence=3	0	50
https://scholarworks.montana.edu/xmlui/bitstream/handle/1/10705/31762100227865.pdf?sequence=1&isAllowed=y	0	32

Figure 2: Metrics for MSU's ScholarWorks IR as shown in the GSC dashboard. The HTTPS root URL is the same as figure 1, but the search type has been changed from "web" to"image."



http://scholarworks.montana...

Performance

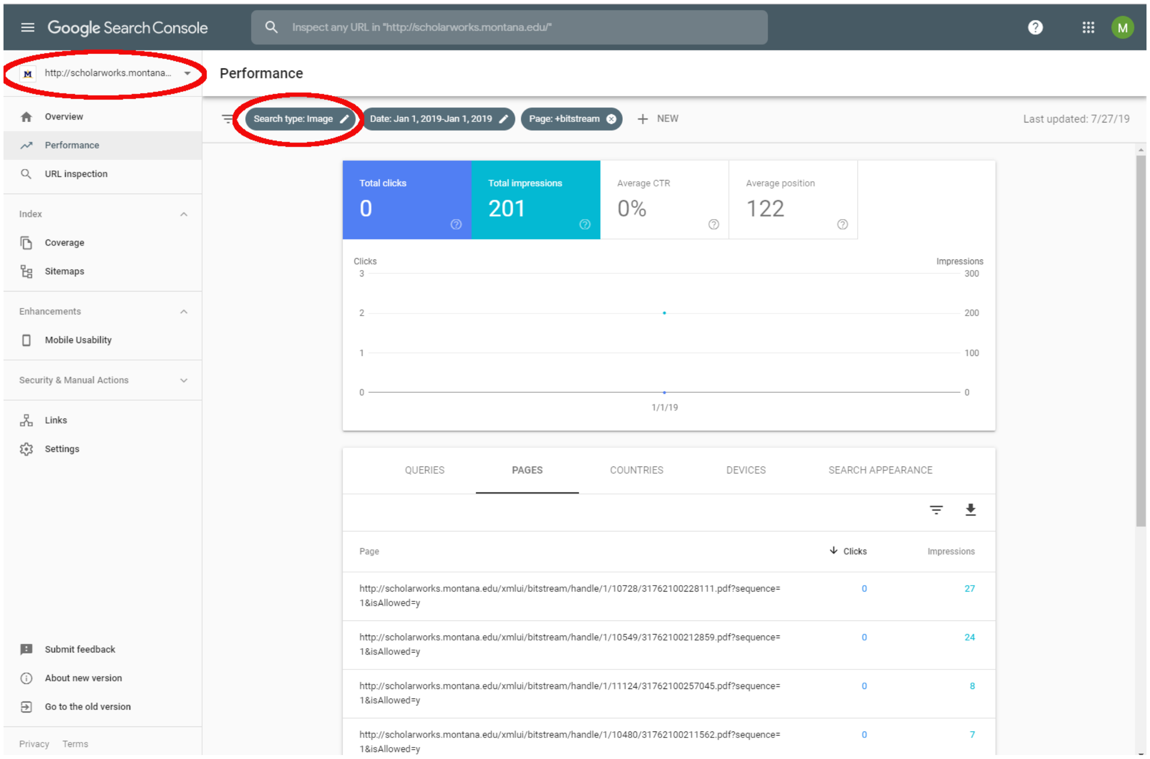
Search type: Web | ate: Jan 1, 2019-Jan 1, 2019 | Page: +bitstream

Last updated: 7/27/19

Total clicks 9 | Total impressions 253 | Average CTR 3.6% | Average position 30.9

Page	Clicks	Impressions
http://scholarworks.montana.edu/xmlui/bitstream/1/5657/1/31762100147568.pdf	2	8
http://scholarworks.montana.edu/xmlui/bitstream/1/6775/1/31762100719598.pdf	2	5
http://scholarworks.montana.edu/xmlui/bitstream/1/7615/1/31762102581939.pdf	1	11
http://scholarworks.montana.edu/xmlui/bitstream/1/5929/1/31762100152493.pdf	1	8
http://scholarworks.montana.edu/xmlui/bitstream/handle/1/8609/31762104283229.pdf	1	8

Figure 3: Metrics for MSU's ScholarWorks IR as shown in the GSC dashboard. The root repository URL is now HTTP instead of HTTPS, and the search type is "web."



http://scholarworks.montana...

Performance

Search type: Image

Date: Jan 1, 2019-Jan 1, 2019

Page: +bitstream

Last updated: 7/27/19

Total clicks 0

Total impressions 201

Average CTR 0%

Average position 122

Page	Clicks	Impressions
http://scholarworks.montana.edu/xmlui/bitstream/handle/1/10728/31762100228111.pdf?sequence=1&isAllowed=y	0	27
http://scholarworks.montana.edu/xmlui/bitstream/handle/1/10549/31762100212859.pdf?sequence=1&isAllowed=y	0	24
http://scholarworks.montana.edu/xmlui/bitstream/handle/1/11124/31762100257045.pdf?sequence=1&isAllowed=y	0	8
http://scholarworks.montana.edu/xmlui/bitstream/handle/1/10480/31762100211562.pdf?sequence=1&isAllowed=y	0	7

Figure 4: Metrics of MSU's ScholarWorks IR as shown in the GSC dashboard. The HTTP root URL is the same as in figure 3, but the search type has been changed once more to "image."

IR root URL	Total clicks from "web" search results	Total clicks from "image" search results
https://scholarworks.montana.edu/	469	2
http://scholarworks.montana.edu/	9	0

Table 1: Total clicks on content files as provided by four separate GSC reports for MSU's ScholarWorks IR for January 1, 2019. The overall total of 480 clicks is automatically calculated by RAMP as shown in Figure 5.

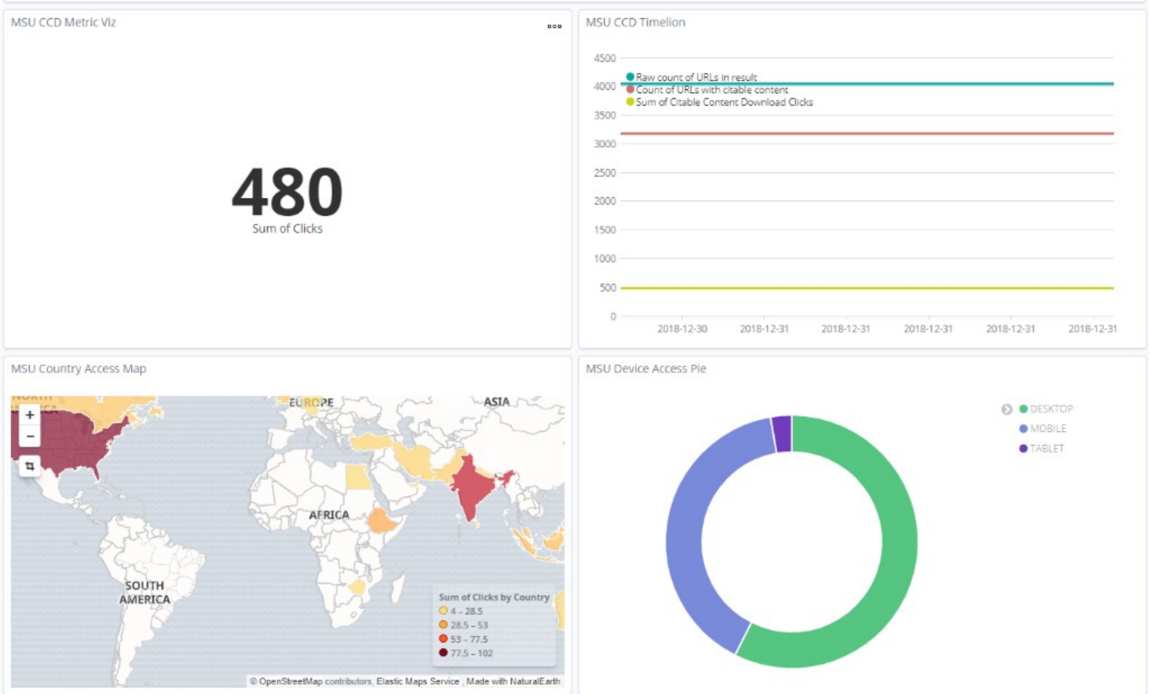


Figure 5: Metrics for MSU's ScholarWorks IR RAMP dashboard for January 1, 2019. Note that the sum of clicks on content files (480) is the same as the sum provided by accessing four different reports in the GSC dashboard. (Please note that the "MSU CCD Timelion" histogram in the top right is level because only one day's worth of data are represented in this dashboard screenshot. There's no apparent trend because data are only collected once per day.)

The manual, multi-step aggregation of separate GSC reports described here is analogous to the RAMP workflow, with the necessary concatenation of data streams performed automatically on ingest of data into RAMP. The remainder of this paper describes the process in detail.

RAMP Data Collection & Publication

Data Description

Data for RAMP are retrieved daily from the GSC API (Google, 2019). Fields and facets[2] listed below are harvested, and "facets" in this case are optional parameters for aggregating the data according to certain values. Harvesting all available data requires downloading two datasets[3]. The first dataset consists of URL or page-level data and will be referred to throughout the remainder of this paper as "page-click" data. The fields and facets harvested for the page-click dataset are:

[2] Please refer to the API documentation for more information about facets.

[3] Prior to August 2018, RAMP data were downloaded as a single dataset that included per-URL SERP performance data with corresponding information about the country from which the search originated and the type of device used. Changes made by Google to the GSC API in summer 2018 impacted RAMP to the extent that including the page facet with the country and device facets resulted in significantly less accurate data. The API update and its effect on RAMP are described in more detail on the project OSF page at https://osf.io/68xpr/. The "files" component includes a report summarizing the GSC API update and the revised RAMP data model.

- **date**: The date the search occurred.
- **page**: The URL of the page that appeared in the Search Engine Results Pages (SERP).
- **position**: The average position of the page in the SERP. Since there are 10 results per page, this number can be divided by 10 to determine, on average, the highest page of the SERP in which the URL appeared.
- **impressions**: The number of times the URL appeared in the SERP.
- **clicks**: The number of clicks on a URL from the SERP.
- **clickthrough**: The ratio of clicks to impressions.

Prior to indexing, page-click data are processed to identify "citable content URLs " that point to non-HTML file types. As reported by (OBrien et al., 2017) citable content URLs are of interest because clicks on these URLs represent access to and potential use of the full content for research, as opposed to views of HTML pages containing abstracts and other metadata. From a standpoint of IR bibliometrics, such use represents potential citations. Thus, URLs that point to content files are considered citable content URLs, and prior to indexing, page-click data are analyzed for strings matching patterns specific to content file URLs for all of the major IR platforms. These patterns, as determined by RAMP administrators, are shown in Table 2, below. Note that whereas the patterns for DSpace and Digital Commons IR are format-independent, the patterns for Fedora and Eprints repositories currently only count citable content URLs that point to PDF files. This limits RAMP's ability to count non-PDF file downloads from Fedora and Eprints.

IR Platform	Content URL Pattern
DSpace	*/bitstream/*
Digital Commons	*/viewcontent.cgi/*
Fedora	*/pdf*
Eprints	*/pdf*

Table 2: URL patterns used by RAMP to identify content file URLs.

Repositories registered with RAMP include a handful of custom-built platforms. Determining the most accurate citable content filters for those repositories is an ongoing process, and for this reason those IR have been excluded from the published data subset.

The identification of citable content URLs results in the addition of a new field to the page-click data:

- **citableContent**: Whether or not a URL that appeared in a SERP points to a content file (PDF, CSV, etc.) Possible values are "Yes" or "No."

Once the page-click data have been processed, they are indexed in Elasticsearch into IR specific page-click indices. That is, each IR in RAMP has its own index of page-click data.

The second dataset includes information about the country of origin and the device used to conduct a search on a Google property that returned IR content in the SERP. These data will be referred to throughout the rest of this paper as "country-device" data. The county-device data overlap somewhat with the page-click data. The fields and facets for this dataset include:

- **date**: The date the search occurred.
- **country**: The country where the search was done. No more granular location information is provided through the API.
- **device**: The type of device used to conduct the search. Values include desktop, mobile, and tablet.
- **position**: Same as above.
- **impressions**: Same as above.
- **clicks**: Same as above.
- **clickthrough**: Same as above.

It is important to note that the country-device data do not include page/URL data. The data are aggregated by combination of country and device, which results a less granular and smaller dataset for any given day when compared to page-click data. Since country-device data cannot be analyzed for citable content URLs, these data are concatenated and indexed immediately after download into IR specific country-device Elasticsearch indices. Each participating IR therefore has two indices in RAMP's ES instance - a page-click index and a country-device index.

It is possible to calculate click sums and other aggregate metrics for both datasets. However, because the data are processed differently, the sum of click counts between the page-click data and country-device data may differ for any given day or time period. Also, the two datasets cannot be meaningfully combined, as there is no shared unique identifier between the datasets that is made available via the API. There is no URL or other page-level identifier in the country-device data, so there is no way within the country-device data to distinguish citable content URLs from HTML URLs.

Output to CSV

The published RAMP data (Wheeler et al., 2019) have been exported from the production Elasticsearch instance and converted to CSV format. The page-click CSV data consist of one row for each page or URL from a specific IR which appeared in search result pages (SERP) within Google properties as described above. The country-device CSV data consist of one row for each unique combination of country and device used to conduct a search for which IR content appeared in the SERP. Also as noted above, daily data are downloaded for each IR in two sets which cannot be combined.

As a result, two CSV datasets are provided for each month of published data:

page-clicks:

The data in these CSV files correspond to the page-level data, and include the following fields:

- **url**: This is returned as a 'page' by the GSC API, and is the URL of the page which was included in an SERP for a Google property.
- **impressions**: The number of times the URL appears within the SERP.
- **clicks**: The number of clicks on a URL which took users to a page outside of the SERP.
- **clickThrough**: Calculated as the number of clicks divided by the number of impressions.
- **position**: The average position of the URL within the SERP.
- **date**: The date of the search.
- **citableContent**: Whether or not the URL points to a content file (ending with pdf, csv, etc.) rather than HTML wrapper pages. Possible values are Yes or No.
- **index**: The Elasticsearch index corresponding to page click data for a single IR. Since the monthly CSV files include data for all participating IR (or all IR included in a subset), index names are needed to extract data for individual IR, or groups of IR.

Filenames for files containing these data end with “page-clicks”. For example, the file named *2019-01_RAMP_subset_page-clicks_v2.csv* contains page level click data for a subset of 35 RAMP participating IR for the month of January, 2019.

country-device-info:

The data in these CSV files correspond to the data aggregated by country from which a search was conducted and the device used. These include the following fields:

- **country**: The country from which the corresponding search originated.
- **device**: The device used for the search.
- **impressions**: The number of times the URL appears within the SERP.
- **clicks**: The number of clicks on a URL which took users to a page outside of the SERP.
- **clickThrough**: Calculated as the number of clicks divided by the number of impressions.
- **position**: The average position of the URL within the SERP.
- **date**: The date of the search.
- **index**: The Elasticsearch index corresponding to country and device access information data for a single IR. Since the monthly CSV files include data for all participating IR (or all IR included in a subset), index names are needed to extract data for individual IR, or groups of IR.

Filenames for files containing these data end with “country-device-info”. For example, the file named *2019-01_RAMP_subset_country-device-info.csv* contains country and device data for all participating IR for the month of January, 2019.

Data Audit

As a final note, RAMP data are sensitive to changes in the GSC API and also to the fact that Google is a "black box" within which unknown data processing and caching methods may impact data harvest and aggregation. These changes can influence RAMP data and specific metrics may vary depending on when data are harvested. For example, because of a previous three-day delay in availability of data through the GSC API, data for May 1, 2019 were not available to RAMP until May 4[4]. The daily download that RAMP uses to index data accounts for this, but due to differences in data caching and resource allocation on Google's end, there is no guarantee that the data for May 1 that RAMP downloads on May 4 will provide the exact same metrics as data for May 1 that stakeholders at participating IR may themselves access from their IR's GSC dashboard.

To assess the accuracy of RAMP data for participants and for the integrity of published research findings based on RAMP data, the RAMP team performs routine, per-day audits of the sum of clicks on URLs pointing to citable content as defined above. This sum, which the RAMP team refers to as “citable content downloads,” or CCD, is the primary metric reported by RAMP. The audit compares daily CCD reported by RAMP with the same information reported by GSC for all of the individual IR in RAMP. For most IR across the full date range, a difference of less than one tenth of one percent is noted. For IR for which a larger difference is noted, the difference is usually the result of a failed or dropped data harvest on one or more days. These audits allow the RAMP team to correct differences where they occur by re-harvesting data for affected IR for specific dates. The RAMP team does not re-harvest data in cases where the difference is less than one tenth of one percent (0.10).

References

Google, Inc. (2020). *Search Console APIs*. Google Developers. https://developers.google.com/webmaster-tools/search-console-api-original/

OBrien, P., Arlitsch, K., Mixter, J., Wheeler, J., & 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*, *35*(1), 144–158. https://doi.org/10.1108/LHT-11-2016-0122

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

Wheeler, Jonathan, Kenning Arlitsch, Minh Pham, Nikolaus Parulian, Patrick OBrien, Jeff Mixter, Montana State University ScholarWorks, University of New Mexico Digital Repository, McMaster University MacSphere, Maryland Shared Open Access Repository (MD SOAR), Digital Repository at the University of Maryland (MD DRUM), Mountain Scholar Digital Collections of Colorado and Wyoming, University of Michigan Deep Blue, Rutgers University RUcore Institutional Repository, Kansas State University Research Exchange (K-REx) , Swarthmore College Works, Bryn Mawr, Haverford, and

[4] A GSC API update in late 2019 reduced this delay to one day. As of this writing, an update to RAMP to account for this change is pending.

Swarthmore Colleges: TriCollege Libraries Institutional Repository, University of Oklahoma, Oklahoma State University, and the University of Central Oklahoma: SHAREOK Repository, University of Nevada Digital Scholarship@UNLV, University of Kentucky UKnowledge, Swedish University of Agricultural Sciences Epsilon Open Archive, Swedish University of Agricultural Sciences Epsilon Archive for Student Projects, Northern Kentucky University Digital Repository, Massey University Massey Research Online, University of Waterloo UWSpace, Caltech Library CaltechAUTHORS Repository, Caltech Library CaltechTHESIS Repository, University of Texas at Austin Texas ScholarWorks, Northeastern University Digital Repository Service, University of Pittsburgh D-Scholarship@Pitt, University of the Western Cape Electronic Theses and Dissertations Repository, University of the Western Cape Research Repository, Royal Roads University and Vancouver Island University VIURRSpace, University of Strathclyde Strathprints, University of Montana ScholarWorks, Virginia Tech VTechWorks, Sam Houston State University Scholarly Works @ SHSU, Indiana University - Purdue University Indianapolis (IUPUI) ScholarWorks, University of Plymouth PEARL, Digital Commons @ The University of Nebraska Lincoln, University of Wollongong Australia: Research Online (2019). RAMP data subset, January 1 through May 31, 2019, University of New Mexico, Dataset, https://doi.org/10.5061/dryad.fbg79cnr0

Wheeler, J., OBrien, P., Mixter, J., & Arlitsch, K. (2019, April 23). *Repository Analytics and Metrics Portal—RAMP*. Open Science Framework. https://osf.io/zx5q7/