RELEASE 5.0.2

FRIENDLY TECHNICAL NOTES FOR PROVIDERS

This is a straightforward technical introduction for publishers and vendors implementing COUNTER-compliant usage statistics.

Tasha Mellins-Cohen

COUNTER

CONTENTS

INTRODUCTION

This is a friendly introduction to the COUNTER Code of Practice Release 5.0.2 for publishers and other providers, and is not intended as a developer's specification manual. It accompanies Release 5.0.2: The Friendly Guide for Providers, which explains the metrics, attributes and reports associated with the Code of Practice.

For more information about implementation please review Release 5.0.2 Appendix D: Guidelines for Implementation at https://cop5.projectcounter.org/.

TRACKING USAGE

Usage data can be generated in several ways and COUNTER does not prescribe which approach should be taken. The two most common approaches are page tagging and log file analysis. Both have advantages and disadvantages, summarized below. It is important to remember that data collected for COUNTER reports only records actual usage: because every platform records usage slightly differently, it is not possible for us to describe specific mechanisms for cleansing the data. This guide therefore only outlines the basic requirements.

PAGE TAGS

Page tags are small pieces of code embedded in each page of your website. They are usually written in JavaScript, though other languages such as Java are used at the discretion of the site developers. Data is gathered via these tags and passed to a database. Scripts written in languages such as JQuery and AJAX can then be used in conjunction with a server-side scripting language such as PHP to manipulate and store the data, allowing complete control over how the data is represented. The data storage and manipulation script may have access to additional information about the web client or the user; for example, by reading information from your access management system.

Page tagging is standard in web analytics (e.g. Google Analytics). One key difference between log file analysis and page tagging is that with page tags a usage count is activated by opening the page in the browser, not by requesting it from the server. This means you are likely to see a more accurate reflection of usage using page tags, because cached pages are counted in the same way as server calls.

Page tags are particularly useful for companies that do not have access to their own web servers; with the increasing use of cloud storage, page tagging is becoming a preferred way to obtain analytics information. Page tagging and tag data analysis can be done in-house, but are also widely available as third-party services.

The table below shows a small sample of page tags from the Google Tag Manager collection. All have direct applications in COUNTER reporting.

Tag name	Google definition	Specific use in Release 5
Page View	The most basic tag, Page View should fire on every page of your site	No specific use
Event	Used to track a specific action or event, such as a button click	Separating out investigations, requests, searches, and other metrics
Timing	Used to track loading speeds on your page	Distinguishing double clicks

If you are interested in using page tagging to generate COUNTER reports, https://analytics.google.com and https://tagmanager.google.com are a good place to start.

LOG FILES

Log files are text files representing individual HTTP requests, including the user host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL and the browser information. HTTP requests are what happen when a user tries to access a web page, while an HTTP response is the delivery of that page to the user's browser.

Most web servers produce log files by default in a pre-defined format which may differ by server type. The log file data you need for your COUNTER implementation should already be available to you without changes to your website.

Because log file data is held on your own servers in a standard format you can use a variety of analytics programmes and receive consistent results over time. Log files are also independent of your users' web browsers, meaning you can reliably track all activity for the purposes of COUNTER reporting.

Be aware that cached pages are not counted in log file analysis because they are not requested from the server, although cached pages can account for a sizeable proportion of page views.

As different servers will deliver log files in different formats, most log file analysis is performed by the people who support the server(s) – usually the vendor's in-house team.

If you are interested in using log file analysis for generating COUNTER reports we recommend you speak to your development team.

If you wish to learn more about log files the Amazon Web Service documentation is well-written and helpful: https://docs.aws.amazon.com/.

COOKIES

Cookies are small files stored on a user's computer and designed to hold a modest amount of data specific to a particular browser accessing a particular website. They can be accessed by either the web server or the user's computer. Page tags can be used to manage the process of assigning cookies to a user's browser; with log file analysis, the server must be configured to do this. There are legal considerations around assigning cookies, so please check the relevant requirements before configuring your setup (e.g. https://gdpr.eu/cookies/).

ABNORMAL SPIKES IN USAGE

What is regarded as an abnormal spike in usage can vary from one institution to another and there are many occasions on which exceptionally high usage in a month is genuine, so we do not have a strict protocol for dealing with usage spikes. The following approaches will provide an indication of possible abnormal usage or another unusual event. These should only be as a prompt for human intervention to take a closer look at the numbers, rather than any automated cut-off of access.

Positive spike in usage

Reported usage may be too high (a positive spike) if, in a specific month, the reported usage by a customer for an individual product is at least one hundred units of measurement greater than 300% above the previous twelve-month average.

Negative spike in usage

Reported usage may be too low (a negative spike) if, in a specific month, the reported usage by a customer for an individual product is less than 1% of the previous twelve-month average, where the average usage of that product in the previous 12 months is at least twenty units of measurement.

SCENARIO

For the period June 2020 to May 2021, Institution Omega users have tracked an average of 100 Total_Item_Requests per month for Title X. In June 2021, the institution's users generate 450 Total_Item_Requests.

The calculation to determine whether this is a positive spike is as follows:

▪ Average usage for the previous 12 months:	100 Total_Item_Requests
▪ 300% above average usage:	300 Total_Item_Requests
▪ Plus 100 units of measurement:	400 Total_Item_Requests

At 450 Total_Item_Requests in June 2021, Institution Omega has shown a positive spike and the publisher of Title X should investigate.

By contrast, Institution Omega users tracked an average of 500 Total_Item_Requests per month for Title Y over the same period, but recorded only 4 Total_Item_Requests in June 2021.

The calculation to determine whether this is a negative spike is as follows:

▪ Average usage for the previous 12 months:	500 Total_Item_Requests
▪ 1% of average usage:	5 Total_Item_Requests

At 4 Total_Item_Requests in June 2021, Institution Omega has shown a negative spike and the publisher of Title Y should investigate.

SEARCHES

A search is counted any time the system executes a search to retrieve a new set of results. This means that systems performing multiple searches (e.g. search for exact match, search for words in subject, general search) to return a single set of results ordered by relevancy must only count a single search, not multiple searches. Design options and activities that do count as separate searches include:

- Bento-box or multi-tab user interfaces, where multiple searches are conducted to retrieve and present multiple result sets.
- Refinement of a set of search results by faceting, where applying a facet or filter requires the search to be re-run.
- Browsing through a topics list or subject authority file, where clicking on the topic or subject conducts a search to present a set of search results.

Note that link resolution never counts as a search.

SCENARIO

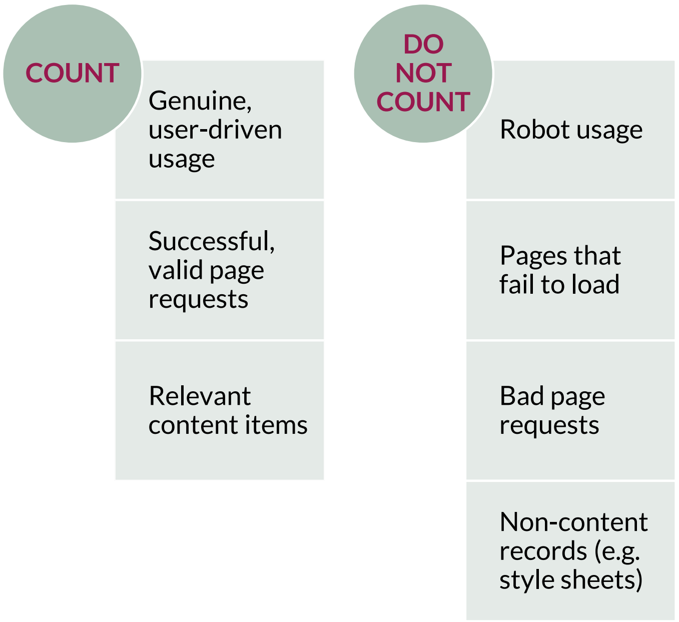
A user clicks on a link in a table of contents alert email. The site resolves the link by 'searching' for the article id, and presents the user with the article full text. No search is counted because this is a link resolution action, but the site should record one count each for Total_Item_Investigation, Unique_Item_Investigation, Total_Item_Request, and Unique_Item_Request

The user proceeds to search for 'history of antibiotics' on the platform, but chooses not to restrict their search to a specific database. This creates one count under each of the Searches_Platform and Searches_Regular metrics. The search returns several hundred results, so the user facets the results list to display only results published since 2017. This creates a second count under each of the Searches_Platform and Searches_Regular metrics.

PROCESSING YOUR DATA

WHAT TO COUNT

COUNTER is designed to provide consistent and credible usage statistics that allow libraries to compare usage across platforms. Release 5 of the Code of Practice therefore outlines what should and should not be counted when creating reports (per the figure opposite).



COUNT

Genuine, user-driven usage

Successful, valid page requests

Relevant content items

DO NOT COUNT

Robot usage

Pages that fail to load

Bad page requests

Non-content records (e.g. style sheets)

RETURN CODES

Return codes are a way to track the success or otherwise of page deliveries (i.e. pages loading successfully in a browser). For web server log files, successful requests are those with specific National Center for Supercomputing Applications (NCSA) return codes, namely all 200s and 304. The standards for return codes are defined and maintained by NCSA, so where your server uses key events their definition should match the NCSA standards. For more information see http://www.ncsa.illinois.edu.

SESSIONS

Release 5 of the Code of Practice includes four ‘Unique’ metrics (Unique_Item_Investigations, Unique_Item_Requests, Unique_Title_Investigations, and Unique_Title_Requests). If a user accesses the full text of a book chapter three times during a single session, this counts as a single Unique_Item_Request. To report on ‘Unique’ metrics, it is necessary to track user sessions.

A user session is typically defined as a logged session ID plus a transaction date. When a session ID is not explicitly tracked, the day should be divided into 24 one-hour slices and a surrogate sessionID generated by combining the transaction date plus the one-hour time slice plus one of: a logged userID; a logged user cookie; or a combination of IP address plus User Agent.

SCENARIO

A platform uses the pattern [IP address] | [User Agent] | [date] | [time]. For a user session with this information:

- Transaction date/time: 2021-06-15 13:35
- IP Address: 192.1.1.168
- User Agent: Mozilla/5.0

The platform will generate a Session ID of 192.1.1.168 | Mozilla/5.0 | 2021-06-15 | 13

DOUBLE CLICKS

Double clicks on an http link should be counted as one request. For the purposes of COUNTER, the time window for a double click is set at a maximum of 30 seconds between the first and second mouse clicks. This means a click at 10:01:00 and a second click at 10:01:29 would be considered a double click, while a click at 10:01:00 and a second click at 10:01:35 would count as two separate single clicks.

A double click may also be triggered by pressing a refresh or back button, and double click filtering applies to all COUNTER Release 5 metrics other than Searches. When two requests are made for the same article within 30 seconds, the first request should be removed and the second retained. Any additional requests for the same article with another 30 seconds should be treated identically: always remove the first and retain the second.

There are several ways to track double-clicking, depending on how the user is authenticated on your site. These options are listed in order of increasing reliability, with option 4 being the most reliable:

1. If the user is authenticated only through their IP address, that IP should be used as the field to trace double clicks. Where you have multiple users on a single IP address, this can lead to separate clicks from different users being logged as a double click from one user, where multiple users are clicking on the same content within a few seconds of each other.
2. When a session cookie is implemented and logged, the session cookie should be used to trace double clicks.
3. When a user cookie is available and logged, the user cookie should be used to trace double clicks.
4. When an individual has logged in with their own profile, their username should be used to trace double clicks.

UNIQUE ITEMS AND TITLES

Unique_Item_Requests and Unique_Item_Investigations count the number of unique items that attracted a certain activity. An item is the unit of content (e.g. articles, book chapters, and multimedia content), and should be identifiable using a unique ID such as a DOI (digital object identifier). If a user requests the same item on more than one occasion in a single session, only one unique activity should be counted for that item.

Similarly, Unique_Title_Investigations and Unique_Title_Requests count the number of unique titles that had a certain activity. A title is the parent entity to which an item belongs, such as a journal or a database. If a user requests multiple items from a single title in a single session, only one unique activity should be counted for that title.

SCENARIO

A user is researching the history of antibiotics. From a list of search results, they open three abstracts and a video record. All four items are different, but two of the abstracts are from chapters in the same book. The counts are:

▪ Total_Item_Investigations:	4
▪ Unique_Item_Investigations:	4
▪ Unique_Title_Investigations:	2

There are two Unique_Title_Investigations because the abstracts come from two different titles, and the multimedia item does not belong to a title.

After reading the abstracts, the user downloads full text PDFs for two chapters, both from the same book. This changes the counts to:

▪ Total_Item_Investigations:	6
▪ Unique_Item_Investigations:	4
▪ Unique_Title_Investigations:	2
▪ Total_Item_Requests:	2
▪ Unique_Item_Requests:	2
▪ Unique_Title_Requests:	1

FEDERATED AND BOT SEARCHES

The growing use of federated searches and the spread of web crawler robots have the potential to inflate usage statistics, so COUNTER requires you to identify this type of usage in your reports. The most common ways to recognize federated and automated search activity are as follows:

- A federated search engine may be using its own dedicated IP address, which can be identified and used to separate out the activity.
- If the standard HTML interface is being used (e.g. for screen scraping), the browser ID within the web log files can be used to identify the activity as coming from a federated search.

- For Z39.50 activity (http://www.niso.org/standards/resources/Z39.50_Resources), authentication is usually through a username/password combination. Create a unique username/password that just the federated search engine will use.
- If an API (application programming interface) gateway is available, set up an instance of the gateway that is for the exclusive use of federated search tools. It is recommended you also require the federated search engine to include an identifying parameter when making requests to the gateway.

COUNTER has lists of federated search tools and web robots in Appendices G and I to the Code of Practice, which are reviewed and updated on a regular basis and which can be found at: https://github.com/atmire/COUNTER-Robots.

Where federated or automated usage is genuine user-driven usage, in the context of Text & Data Mining, the Access_Method 'TDM' attribute should be used. This allows users of the resultant reports to distinguish automated usage from more traditional ('regular') usage.

PROTOCOL BULK DOWNLOAD TOOLS

Usage of full text articles initiated by automatic or semi-automatic bulk download tools, such as Quosa or Pubget, should only be recorded when the user has clicked on the downloaded full text article in order to open it.

DATA CHANGES

RETROSPECTIVE REPORTING OF ERRORS IN USAGE DATA

If errors are identified in COUNTER reports, the provider must correct the errors within three months of their discovery and inform their customers of the corrections.

REPORTING OF USAGE STATISTICS WHEN JOURNAL TITLES CHANGE

When the title of a journal is changed but the identification code (ISSN or DOI) stays the same, you should consider the journal to be a single 'Title' for the purposes of COUNTER reporting. Reports should be provided against the new title, with the original title being dropped from the list.

If a new DOI or ISSN is allocated to the journal when the title changes, you should consider the journal to be two 'Titles' and provide usage information separately. You must not report usage for the same period under both the old and new DOI or ISSN. Any report generated for the old DOI or ISSN should show zero usage from the month in which the new DOI or ISSN takes effect.

DELIVERING COUNTER REPORTS

COUNTER reports are available in two formats: the primary format is JSON, delivered through SUSHI, with delimited (spreadsheet) files as a secondary format. There are some key factors to consider in delivering COUNTER reports:

- Reports must be provided monthly.
- Data must be updated within four weeks of the end of the reporting period.
- A minimum of the current calendar year's data plus 24 months of back data must be available, unless you are newly COUNTER compliant (i.e. on 30 July 2021, the data for 2019, 2020, and January to June 2021 must be available).
- It must be possible to request usage for a date range in months, within the most recent 24-month period.
- Where no date range is specified, the default is to show the whole of the most recent 24-month period.
- Reports should be readily available on a password-controlled website.
- There should be an option to receive an email alert when a new report is available.
- Each report should reside in a separate file or page to avoid producing files of unwieldy size.
- Usage statistics reported in the COUNTER reports must not be browser-dependent, and you must support current versions, compliant with World Wide Web Consortium (WC3) standards, of the following web browsers: Google Chrome, Microsoft Edge, and Mozilla Firefox.

Publishers must provide COUNTER reports on a per-customer ID basis. That is, if a business school has a separate customer ID from its parent university, the school and the university should be sent separate COUNTER reports. This applies whether authentication is through IP address recognition, Shibboleth, or other mechanisms. To follow the example above, if the business school shares the parent university's IP range and relies on IP recognition for authentication, it will not be possible to distinguish usage from the school from that of the university and therefore only the university should receive a COUNTER report.

DELIMITED FILES

The reports specified in COUNTER Release 5 can all be delivered as delimited files in either comma separated (.csv) or tab separated (.tsv) format. Delimited files can be opened and read in all spreadsheet tools, including Excel, OpenOffice Calc, Google Sheets, and Numbers for Mac Formatting in the sense of typeface and colour, are irrelevant in delimited files, but adherence to the layout described in the COUNTER specification for each report is required for compliance.

SUSHI

The Standardized Usage Statistics Harvesting Initiative (SUSHI) protocol is designed to simplify the gathering of usage statistics by librarians, and SUSHI support is mandatory for compliance with COUNTER Release 5. There are four API paths that must be supported for Release 5, as shown below. More information on the COUNTER SUSHI specification can be found on our website at https://www.projectcounter.org/counter-sushi/.

GET/status

- Returns the current status of the COUNTER_SUSHI API service

GET/reports

- Returns a list of reports supported by the COUNTER_SUSHI API service

GET/reports/{ReportID}

- Returns a specific supported report, such as GET/reports/TR_B1 for Book Requests (Excluding OA_Gold)

GET/members

- Returns the list of consortium members or sites for multi-site customers

AUDITING

An important feature of the COUNTER Code of Practice is that compliant providers must be independently audited on a regular basis to maintain their COUNTER-compliant status. The audit is designed to meet the need of libraries for credible usage statistics without being too onerous for publishers or providers. For this reason, audits are conducted online using the process outlined in the auditing standards and procedures described in Appendix E of the Code of Practice at https://cop5.projectcounter.org/.

An independent audit is required within six months of first achieving COUNTER compliance, and annually thereafter. COUNTER will recognize an audit carried out by any Certified Public Accountant in the USA, by any Chartered Accountant in the UK, or by their equivalent in other countries. Alternatively, the audit may be conducted by one of our COUNTER-approved auditors, ABC (https://www.abc.org.uk/) and BPA Worldwide (https://bpaww.com/).

THE COUNTER VALIDATION TOOL

To help your audit go smoothly, start by checking your reports with the COUNTER Validation Tool (available at https://stats.redi-bw.de/counter-r5-validation-preview/). The Tool detects missing metrics and similar issues, for example a report including fewer Total_Item_Requests than Unique_Item_Requests for a title, which would be a critical error. Other common errors include missing Master Reports and creating Standard Views independently of the Master Reports.

THE AUDIT PROCESS

COUNTER-compliant vendors are notified in writing by COUNTER at least three months before an audit is due. You have one month to respond to the notice, telling us your planned timetable for the audit and the name of the organization that will carry it out, as well as sharing any queries you have about the audit process.

Before your audit begins, we recommend arranging a meeting with the auditor to show them around your platform and to flag up any areas of concern. A briefing document for the platform, describing your mechanism for tracking COUNTER metrics and so forth, is also helpful. Regardless of the auditor selected the audit must adhere to the requirements and use the tests specified in Appendix E of the COUNTER Code of Practice, covering the format and structure of the reports, the integrity of the reported statistics, and delivery of the reports.

On completion of a successful audit the auditor must send a signed copy of the audit report to the COUNTER office (lorraine.estelle@counterusage.org). If the audit is not successful, the auditor must send an interim report to the COUNTER office outlining the reasons for failure. The auditors will then work with you to correct the areas of failure within a timeframe agreed to by COUNTER.

CATEGORIES OF AUDIT RESULT

Pass

No further action is required following the audit. In some cases the auditor may add observations to the audit report, designed to help improve your COUNTER usage reports, but they are not requirements for compliance.

When you pass a COUNTER audit you will be provided with an updated logo confirming your compliance, which you may wish to add to your website and link back to your entry on the COUNTER Registry of Compliant Publishers. COUNTER will also contact you to confirm that the details held in the Registry are correct.

Even after a successful audit, things can go awry, so we recommend you regularly use the free COUNTER Validation Tool to test both tabular and SUSHI reports.

Qualified Pass

The audit has revealed a minor issue which needs to be addressed. A minor issue does not affect the reported figures (it may be related to the presentation of the report, for example), but needs to be resolved within three months of the audit to maintain COUNTER-compliant status.

Fail

There is an issue that must be rectified to maintain COUNTER-compliant status. You will be given a grace period of one month from the date of notification to rectify the reasons for the failure and achieve a pass.

ABOUT THE AUTHOR

After two decades in scholarly publishing Tasha is now an independent publishing consultant. Having worked with society and commercial publishers she has a deep appreciation for the changing pressures on publishers, funders, researchers and research institutions and uses that to partner with publishers to develop data-driven business models that will allow them to achieve a sustainable transition to Open Access.

She is an active participant in the scholarly publishing community as a member of the COUNTER Executive Committee, and regularly volunteers time to Jisc, UKSG, OASPA, and other industry bodies. She can be found on LinkedIn at www.linkedin.com/in/tashamc.

COUNTER

This guide is published by COUNTER in 2021
25 Egbert Road, Winchester, SO23 7EB
lorraine.estelle@counterusage.org