THE FRIENDLY GUIDE TO RELEASE 5

TECHNICAL NOTES FOR PROVIDERS

Tasha Mellins-Cohen

COUNTER

CONTENTS

INTRODUCTION

This Guide is a friendly introduction to COUNTER implementation for publishers and other providers, and is not intended as a developer's specification manual. It accompanies *The Friendly Guide to Release 5 for Providers*, which explains the metrics, attributes, and reports associated with COUNTER Release 5.

For more information about implementation, please read the COUNTER Code of Practice, Release 5, Appendix D: Guidelines for implementation.

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 cleaning up the data. This Guide therefore only outlines the 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 that 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.

Page tag examples

This is a small sample of page tags from the Google Tag Manager collection, all of which have direct applications in COUNTER reporting.

Tag Name	Google definition	Specific Use in Release 5 Master Reports
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 for generating COUNTER reports, www.google.co.uk/analytics/ is 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 users browser.

Most web servers produce log files by default, in a pre-defined format which may differ by server type, so the log file data you need for your COUNTER implementation should already be available to you, without your having to make 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 significant 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 that you speak to your development team.

To learn more about log files generally, the Amazon Web Service log file documentation at aws.amazon.com/documentation is well-written and helpful.

Cookies

Cookies are small files that are stored on a user's computer. They are designed to hold a modest amount of data specific to a particular browser accessing a particular website, and 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 requirements that apply to you before configuring your setup.

ABNORMAL SPIKES IN USAGE

What is regarded as an abnormal spike in usage can vary from one institution to another; 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. We suggest that 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.

SCENARIO

For the period June 2017 to May 2018, Camford users have tracked an average of 100 Total_Item_Requests per month for Title X.

In June 2018, Camford users generate 450 Total_Item_Requests: this is a positive spike in usage.

The calculation to trigger the 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

Negative spike in usage

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

SCENARIO

For the period June 2017 to May 2018, Camford users have tracked at 500 Total_Item_Requests per month for Title X.

In June 2018, Camford users generate 4 Total_Item_Requests: this is a negative spike in usage.

The calculation to trigger the positive spike is as follows:

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

SEARCHES

A search is counted any time the system executes a search to retrieve a new set of results. This means that systems that perform 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.

Things that do count as separate searches:

- 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

Susan clicks the following link after receiving an email:
http://www.pubalpha.org/showArticle?id=12345

The site resolves the link by 'searching' for id=12345

No search is counted, because this is a link resolution action

Having read the article, Susan then searches for 'history of antibiotics' across pubalpha.org, creating one count under the Searches_Platform metric. The search returns several thousand results, so she facets the list to show only results from journals; this creates a second count under the Searches_Platform metric.

PROCESSING YOUR DATA

WHAT TO COUNT

- Genuine, user-driven usage
- Successful, valid page requests: pages that fail to load, or bad requests, must be excluded
- Relevant content items: other records (e.g. style sheets) must be excluded

WHAT NOT TO 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—that is, did this page load successfully on the user's 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. Where your server uses key events, their definition should match the NCSA standards. For more information, see www.ncsa.illinois.edu.

SESSIONS

Release 5 of the Code of Practice includes four 'Unique' metrics. For example, if a user accesses the fulltext of a book chapter three times during a single session, this counts as a single Unique_Item_Request. In order to report on 'Unique' metrics, it is therefore 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 the following:

- A logged userID
- A logged user cookie
- A combination of IP address plus User Agent

SCENARIO

Consider the following transaction:

- transaction date/time: 2017-06-15 13:35
- IP Address: 192.1.1.168
- User Agent: Mozilla/5.0

This allows us to generate a Session ID of 192.1.1.168 | Mozilla/5.0 | 2017-06-15 | 13 following the pattern [IP address] | [User Agent] | [date] | [time]

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. For example, a click at 10.01.00 and a second click at 10.01.29 would be considered a double click; 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. When two requests are made for the same article within 30 seconds for PDF, 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 different 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 occasionally lead to separate clicks from different users being logged as a double click from one user. This will only happen if the multiple users are clicking on exactly 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

Two COUNTER metric types—Unique_Item_Requests and Unique_Item_Investigations—count the number of unique items that had a certain activity. An item is the typical unit of content being accessed by users, such as articles, book chapters, and multimedia content; these should be identifiable using a unique ID such as a DOI (digital object identifier). A title is the parent entity to which an item belongs—for example, a journal or a database.

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, two COUNTER metrics types—Unique_Title_Investigations and Unique_Title_Requests—count the number of unique titles that had a certain activity. A title would usually be something like a book or journal—the wrapper around items of content.

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

Susan is researching the history of antibiotics on Publisher Platform Alpha. From a list of search results, she opens three article abstracts and a video record. All four records are different, but two of the articles are from the same journal. The counts are:

- Total_Item_Investigations: 4
- Unique_Item_Investigations: 4
- Unique_Title_Investigations: 3

After reading the abstracts, Susan downloads the PDFs for two of the articles, both from the same journal. The counts change to:

- Total_Item_Investigations: 6
- Unique_Item_Investigations: 4
- Unique_Title_Investigations: 3
- 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 (described at 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 that you also require the federated search to include an identifying parameter when making requests to the gateway.

COUNTER has lists of federated search tools and web robots in the appendices to the full implementation guide, which are reviewed and updated on a regular basis.

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 fulltext articles that is 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 fulltext 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 DOI or ISSN and the new. 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 year's data plus 24 months of back data must be available, unless you are newly COUNTER compliant (i.e. on 10 July 2018, the data for 2016, 2017, and Jan-June 2018 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, Internet Explorer and Mozilla Firefox.

Publishers must provide COUNTER reports on a per-customer ID basis. For example, if a business school has a separate customer ID from its parent university, the school and the university should be sent separate COUNTER reports. Most authentication is through IP address recognition. In the example above, if the business school does not have a unique IP range, it is not possible to distinguish usage from the school from that of the university, and therefore only the university should receive a COUNTER report.

SCENARIO

Camford and the Camford School of Business subscribe to different products on Publisher Platform Alpha. They share an IP range, and Publisher Platform Alpha cannot distinguish between the institutions for reporting purposes: Camford's COUNTER reports include all of the usage from the Camford School of Business.

Oxbridge and the Oxbridge School of Business also subscribe to different products on Publisher Platform Alpha. They have distinct IP ranges, so Publisher Platform Alpha can distinguish between the institutions for reporting purposes: Oxbridge's COUNTER reports exclude all of the usage from the Oxbridge School of Business.

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:

- 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

More details on the COUNTER SUSHI specification can be found on the COUNTER website at www.projectcounter.org/COUNTER_SUSHI5_0.html.

DELIMITED FILES

The reports specified in COUNTER Release 5 can all be delivered as delimited files:

- Comma separated, or .csv
- Tab separated, or .tsv

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.

AUDITING

An important feature of the COUNTER Code of Practice is that compliant providers must be independently audited on a regular basis in order to maintain their COUNTER-compliant status. We have tried to ensure that the audit meets the need of libraries for credible usage statistics without making the process too onerous. For this reason, audits are conducted online using the process outlined in the auditing standards and procedures (Appendix E of the Code of Practice).

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 done by one of our COUNTER-approved auditors.

THE AUDIT PROCESS

COUNTER-compliant vendors are notified in writing by COUNTER when an audit is required. We send this notification at least three months before the audit is due.

You have one month to respond to the notice, telling us:

- Your planned timetable for the audit
- The name of the organization that will carry out the audit
- Any queries you have about the audit process

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. The audit is carried out in three stages:

1. The format and structure of the usage reports
2. The integrity of the reported usage statistics
3. The delivery of the usage 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 will send an interim report to the COUNTER office outlining the reasons for failure. The auditors will 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 as a result of the audit. In some cases the auditor may add observations to the audit report. These are designed to help improve your COUNTER usage reports, but they are not requirements for compliance.

Qualified Pass

The audit has been passed, but with a minor issue which needs to be addressed in order to maintain COUNTER-compliant status. A minor issue does not affect the reported figures; for example, it may be related to the presentation of the report. Minor issues need to be resolved within three months of the audit to maintain COUNTER-compliant status.

Fail

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