COUNTER Code of Practice

7.0 Processing Rules for Underlying COUNTER Reporting Data

Usage data collected by content providers for the usage reports to be sent to customers should meet the basic requirement that only intended usage is recorded and that all requests that are not intended by the user are removed.

Because the way usage records are generated can differ across platforms, it is impractical to describe all the possible filters and techniques used to clean up the data. This Code of Practice, therefore, specifies only the requirements to be met by the data to be used for building the usage reports.

7.1 Return Codes

Only successful and valid requests MUST be counted. For web server log files successful requests are those with specific W3C Status Codes, codes (200 and 304). The standards for return codes are defined and maintained by W3C (http://www.w3.org/Protocols/HTTP/HTRESP.html). If key events are used, their definition MUST match the W3C standards. (For more information see The Friendly Guide to Release 5: Technical Notes for Content Providers.)

7.2 Double-Click Filtering

The intent of double-click filtering is to remove the potential of over-counting which could occur when a user clicks the same link multiple times, typically due to a slow internet connection. Double-click filtering applies to Total_Item_Investigations, Total_Item_Requests, No_License and Limit_Exceeded. See Sections 7.3 and 7.4 below for information about counting unique items and titles. The double-click filtering rule is as follows:

Double-clicks, i.e. two clicks in succession, on a link by the same user within a 30-second period MUST be counted as one action. For the purposes of COUNTER, the time window for a double-click on any page 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 (one action); a click at 10:01:00 and a second click at 10:01:35 would count as two separate single clicks (two actions).

A double-click may be triggered by a mouse-click or by pressing a refresh or back button. When two actions are made for the same URL within 30 seconds the first request MUST be removed and the second retained.

Any additional requests for the same URL within 30 seconds (between clicks) MUST be treated identically: always remove the first and retain the second.

There are different ways to track whether two requests for the same URL are from the same user and session. These options are listed in order of increasing reliability, with Option 4 being the most reliable.

1. If the user is authenticated only through an IP address, that IP address combined with the browser's user-agent (logged in the HTTP header) MUST be used to trace double-clicks. Where you have multiple users on a single IP address with the same browser user-agent, 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 MUST be used to trace double-clicks.
3. When a user cookie is available and logged, the user cookie MUST be used to trace double-clicks.
4. When an individual has logged in with their own profile, their username MUST be used to trace double-clicks.

7.3 Counting Unique Items

Some COUNTER Metric_Types count the number of unique items that had a certain activity, such as a Unique_Item_Requests or Unique_Item_Investigations.

For the purpose of COUNTER metrics, an item is the typical unit of content being accessed by users, such as articles, book chapters, book sections, whole books (if delivered as a single file), and multimedia content. The item MUST be identified using the unique ID which identifies the work (e.g. chapter or article) regardless of format (e.g. PDF, HTML, or EPUB). If no item-level identifier is available, then use the item name in combination with the identifier of the parent item (i.e. the article title + ISSN of the journal, or chapter name + ISBN of the book).

The rules for calculating the unique item counts are as follows:

If multiple transactions qualifying for the Metric_Type in question represent the same item and occur in the same user-sessions, only one unique activity MUST be counted for that item.

A user session is defined any of the following ways: by a logged session ID + transaction date, by a logged user ID (if users log in with personal accounts) + transaction date + hour of day (day is divided into 24 one-hour slices), by a logged user cookie + transaction date + hour of day, or by a combination of IP address + user agent + transaction date + hour of day.

To allow for simplicity in calculating session IDs, when a session ID is not explicitly tracked, the day will be divided into 24 one-hour slices and a surrogate session ID will be generated by combining the transaction date + hour time slice + one of the following: user ID, cookie ID, or IP address + user agent. For example, consider the following transaction:

- Transaction date/time: 2017-06-15 13:35
- IP address: 192.1.1.168
- User agent: Mozilla/5.0
- Generated session ID: 192.1.1.168|Mozilla/5.0|2017-06-15|13

The above replacement for a session ID does not provide an exact analogy to a session. However, statistical studies show that the result of using such a surrogate for a session ID results in unique counts are within 1–2 % of unique counts generated with actual sessions.

7.4 Counting Unique Titles

Some COUNTER Metric_Types count the number of unique titles that had a certain activity, such as a Unique_Title_Requests or Unique_Title_Investigations.

For the purpose of COUNTER metrics, a title represents the parent work that the item is part of. When the item is a chapter or section, the title is the book. The title MUST be identified using a unique identifier (e.g. an ISBN for a book) regardless of format (e.g. PDF or HTML).

The rules for calculating the unique title counts are as follows:

If multiple transactions qualifying for the Metric_Type in question represent the same title and occur in the same user-session only one unique activity MUST be counted for that title.

A user session is defined any of the following ways: by a logged session ID + transaction date, by a logged user ID (if users log in with personal accounts) + transaction date + hour of day (day is divided into 24 one-hour slices), by a logged user cookie + transaction date + hour of day, or by a combination of IP address + user agent + transaction date + hour of day.

To allow for simplicity in calculating session IDs, when a session ID is not explicitly tracked, the day will be divided into 24 one-hour slices and a surrogate session ID will be generated by combining the transaction date + hour time slice + one of the following: user ID, cookie ID, or IP address + user agent. For example, consider the following transaction:

- Transaction date/time: 2017-06-15 13:35
- IP address: 192.1.1.168
- User agent: Mozilla/5.0
- Generated session ID: 192.1.1.168|Mozilla/5.0|2017-06-15|13

The above replacement for a session ID does not provide an exact analogy to a session. However, statistical studies show that the result of using such a surrogate for a session ID results in unique counts are within 1–2 % of unique counts generated with actual sessions.

7.5 Attributing Usage when Item Appears in More Than One Database

Content providers that offer databases where a given content item (e.g. an article) is included in multiple databases MUST attribute the Investigations and Requests metrics to just one database. The following recommendations may be helpful when choosing when ambiguity arises:

- Give priority to databases that the institution has rights to access.
- If there is a priority order for databases for search or display within the platform, credit usage to the highest priority database.
- Beyond that, use a consistent method of prioritizing database, such as by database ID or name.
- If none of the above, pick randomly.

7.6 Federated Searches

Search activity generated by federated search engines MUST be categorized separately from searches conducted by users on the host platform.

Any searches generated from a federated search system MUST be included in the separate Searches_Federated counts within Database Reports and MUST NOT be included in the Searches_Regular or Searches_Automated counts.

The most common ways to recognize federated 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 user agent within the web log files can be used to identify the activity as coming from a federated search.
- For Z39.50 activity, authentication is usually through a username/password combination. Create a unique username/password that just the federated search engine will use.
- If an API or XML 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 provides lists of user agents that represent the most common federated search tools. See Appendix G.

7.7 Discovery Services and Other Multiple-Database Searches

Search activity generated by discovery services and other systems where multiple databases not explicitly selected by the end user are searched simultaneously MUST be counted as Searches_Automated on Database Reports. Such searches MUST be included on the Platform Reports as Searches_Platform, but only as a single search regardless of the number of databases searched.

Example: A user searches a content site where the librarian has pre-selected 20 databases for business and economics searches. For each search conducted by the user:

- In the Database Report, each of the 20 databases gets credit for 1 Searches_Automated.
- In the Platform Report, Searches_Platform gets credited by 1.

7.8 Internet Robots and Crawlers

Activity generated by internet robots and crawlers MUST be excluded from all COUNTER usage reports. COUNTER provides a list of user agent values that represent the crawlers and robots that MUST be excluded. Any transaction with a user agent matching one on the list MUST NOT be included in COUNTER reports.

COUNTER maintains the current list of internet robots and crawlers at https://github.com/atmire/COUNTER-Robots

7.9 Tools and Features that Enable Bulk Downloading

Only genuine, user-driven usage MUST be reported. COUNTER reports MUST NOT include usage that represents requests of full-text content when it is initiated by automatic or semi-automatic bulk download tools where the downloads occur without direct user action.

- Products like Quosa or Pubget MUST only be recorded only when the user has clicked on the downloaded full-text article in order to open it.
- Full text retrieved by automated processes such as reference manager software or robots (see Section 7.8 above) MUST be excluded.
- Usage that occurs through emailing of a list of articles (Requests) or citations (Investigations) that was not as a result of a user explicitly selecting the items for sharing MUST be excluded. Note that the act of a user explicitly sharing an item would be considered an Investigation, and a user downloading and then emailing a PDF would also be considered a Request.

7.10 Text and Data Mining

Text and data mining (TDM) is a computational process whereby text or datasets are crawled by software that recognizes entities, relationships, and actions. (STM Statement on Text and Data Mining)

TDM does NOT include straightforward information retrieval, straightforward information extraction, abstracting and summarising activity, automated translation, or summarising query-response systems.

A key feature of TDM is the discovery of unknown associations based on categories that will be revealed as a result of computational and linguistic analytical tools.

Principles for reporting usage:

- COUNTER does not record TDM itself, as most of this activity takes place after an article has been downloaded. All we can do is track the count of articles downloaded for the purposes of mining.
- Usage associated with TDM activity (e.g. articles downloaded for the purpose of TDM) MUST be tracked by assigning an Access_Method of TDM.
- Usage associated with TDM activity MUST be reported using the Title, Database, and Platform Master Reports by identifying such usage as Access_Method=TDM.
- Usage associated with TDM activity MUST NOT be reported in Standard Views (TR_J1, TR_B1, etc.).

Detecting activity related to TDM:
TDM activity typically requires a prior agreement between the content provider and the individual or organization downloading the content for the purpose of text mining. The content provider can isolate TDM-related traffic using techniques like:

- Providing a dedicated end-point that is specifically for TDM data harvesting.
- Requiring the use of a special account or profile for TDM data harvesting.
- Assigning an API key that would be used for the harvesting.
- Registering the IP address of the machine harvesting content.

Harvesting of content for TDM without permission or without using the endpoint or protocol supplied by the content provider MUST be treated as robot or crawler traffic and MUST be excluded from all COUNTER reports.