УДК 026.06
DOI 10.33186/1027-3689-2019-12-25-33

М. В. Гончаров, К. А. Колосов
ГПНТБ России

Использование статистических данных веб-серверов библиотек для вычисления альтметрик

Рассмотрены вопросы использования файлов, собираемых веб-серверами библиотек (лог-файлов), в качестве ещё одного источника для вычисления альтметрик. Поскольку лог-файлы содержат информацию о всех запросах к веб-серверу, их можно использовать для вычисления статистических показателей активности удалённых пользователей в соответствии с требованиями ГОСТа 7.0.20-2014. Проанализированы положения этого стандарта в части учёта обслуживания удалённых пользователей. Международным форматом, определяющим перечень и форму предоставления статистических данных по использованию электронных ресурсов, является формат COUNTER. Представлены особенности текущей версии 5 Свода правил COUNTER, правила обработки исходных данных (лог-файлов). Сделан вывод о том, что в настоящее время в библиотеках России нет единой методики подсчёта обращений к электронным ресурсам, которая, с одной стороны, соответствовала бы требованиям ГОСТов, а с другой – поддерживала совместимость с международными стандартами, такими как COUNTER.

Разумное сочетание требований ГОСТа, базовых показателей COUNTER, а также расчётов дополнительных показателей, полученных на основе анализа лог-файлов, позволяет анализировать востребованность электронных ресурсов библиотек различными категориями пользователей.

Статья подготовлена в рамках Государственного задания ГПНТБ России на 2019 г.

Ключевые слова: библиометрия, альтметрики, COUNTER, библиотечная статистика.

UDC 026.06
DOI 10.33186/1027-3689-2019-12-25-33

Mikhail Goncharov and Kirill Kolosov
Russian National Public Library for Science and Technology, Moscow, Russia

Using www-server statistical data for calculating altmetrics

The authors discuss the use of files being accumulated by library www-serves (log-files) for calculating altmetrics. The log-files comprise information on all queries therefore they can be used for calculating statistical indicators of online user activity index in compliance with the federal standard GOST 7.0.20-2014. The authors also analyze the provisions of the standard as referred to online user services control. The COUNTER format is the international format assigning the list and form for presenting statistical data on using digital resources. The current version 5, the COUNTER Code and the algorithm of master data (log-files) processing are discussed. The authors conclude that Russian libraries have not introduced the unified method for counting the requests to their digital resources so far. Meaningful combination of GOST requirements, COUNTER fundamental indicators and calculation of indicators obtained through log-files analysis enable to analyze the relevance of library digital resources to various user categories.

The paper is prepared within the framework of the State Order of RNPLS&T's for the year 2019.

Keywords: bibliometrics, almetrics, COUNTER, library statistics.

The growing trend of accessing the electronic resources of remote user libraries, as well as the intensive increase in content available in electronic format, lead to the accumulation in libraries of significant volumes of files containing the history of user requests (the so-called log files). These files can be used to obtain statistics on the demand for electronic resources, counting remote users, and calculating the use of electronic resources on the Internet, i.e. they are alternative measures of scientific activity, or altmetrics. Altmetric studies cover the entire audience of Internet users, which consists not only of scientists publishing articles and citing the works of their colleagues, but also of those who are outside the scholar community they do not write scientific papers and, accordingly, not engaged in citation. Given the trend of a gradual expansion of the range of altmetric sources, the data generated by libraries based on processing information from log files can be promising source of almetric calculations. Processing of log files collected by the library's web servers should, firstly, calculate the main indicators taken into account in the library statistics in accordance with the requirements of GOST R 7.0.20-2014 "Library statistics: indicators and calculation units", and secondly, the selection of data that allows the calculation of user requests for electronic resources in accordance with the recommendations of international stand-

ards. For the time being in the Russian libraries there is no single methodology for calculating calls to electronic resources, which, on the one hand, would meet the requirements of GOSTs, and on the other, maintain compatibility with international standards, such as COUNTER. At the same time, developers of library automation systems have every opportunity for the gradual practical implementation of such calculations in their software products. A reasonable combination of GOST requirements, COUNTER basic indicators, as well as calculations of additional indicators obtained on the basis of analysis of library web server log files provide rich opportunities for analyzing the demand for library electronic resources by various categories of users and are another significant source of data when calculating altmetrics.

Тенденция роста обращений к электронным ресурсам библиотек удалённых пользователей, а также интенсивное увеличение контента, доступного в электронном формате, приводят к накоплению в библиотеках значительных объёмов файлов, содержащих историю пользовательских запросов (так называемых лог-файлов). Эти файлы можно использовать для получения статистики о востребованности электронных ресурсов [**1**], подсчёта удалённых пользователей, а также для расчёта показателей использования электронных ресурсов в интернете [**2**–**4**], т.е. они являются альтернативными измерителями научной деятельности, или альтметриками (*altmetrics*). Как отмечено в [**5**], альтметрические исследования охватывают всю аудиторию интернет-пользователей, которая состоит не только из учёных, публикующих статьи и ссылающихся на труды своих коллег, но и из тех, кто находится за пределами научного сообщества – не пишет научных работ и соответственно не занимается научным цитированием.

Учитывая тенденцию постепенного расширения круга источников альтметрик, данные, формируемые библиотеками на основе обработки информации из лог-файлов, могут стать ещё одним перспективным источником альтметрических расчётов.

Обработка лог-файлов, собираемых веб-серверами библиотеки, должна обеспечить, во-первых, подсчёт основных показателей, учитываемых в библиотечной статистике в соответствии с требованиями ГОСТа Р 7.0.20-2014 «Библиотечная статистика: показатели и единицы исчисления» [**6**], а во-вторых – отбор данных, позволяющих производить подсчёт обращений пользователей к электронным ресурсам в соответствии с рекомендациями международных стандартов.

Подсчёт пользователей библиотеки. В пункте 7.1.3 ГОСТа 7.0.20-2014 прописано, что количественные показатели посещений библиотеки и обращений пользователей к её электронным ресурсам подсчитываются дифференцированно по целям посещения. Единица подсчёта – приход пользователя в библиотеку или обращение (сессия) к веб-сайту. Под сессией понимается обращение пользователя с одного и того же IP-адреса в течение определённого (фиксированного) времени. При этом количество произведённых обращений (запросов) в течение сессии для расчёта этого показателя не принимается во внимание. Просмотр не менее одной веб-страницы приравнивается к посещению библиотеки.

Подсчёт запросов пользователей. Согласно п. 7.2.1 ГОСТа 7.0.20-2014, самостоятельное обращение пользователя к ресурсам библиотеки запросом не является, подсчёт ведётся в соответствии с п. 7.1.3, т.е. при обработке удалённых запросов учитываются только сессии. Однако в п. 8.1 регламентирован порядок подсчёта выданных из библиотечного фонда документов, в том числе и электронных, а именно – количества выданных/выгруженных электронных документов (в названиях и страницах).

В п. 8.2.2 указано, что единицей подсчёта количества обращений к электронному каталогу и справочно-библиографическим базам является выгруженная запись.

Если руководствоваться положениями ГОСТа 7.0.20-2014, то при подсчёте удалённых посещений библиотеки (на основе сессий) раздельный учёт обращений физических лиц, организаций, роботов, систем сбора данных интернет-поисковиков и т.д. не производится. Как отмечено в [**7**], расчёт количества пользователей по методике ГОСТа возможен только частично. Кроме того, в стандарте не сказано о возможности отслеживать или передавать статистические данные, такие как количество выданных/выгруженных электронных документов (в названиях и страницах), внешним агрегаторам для расчёта альтметрик.

Такую возможность предусматривает ГОСТ 57723-2017 «Информационно-коммуникационные технологии в образовании. Системы электронно-библиотечные. Общие положения» [**8**]. В п. 4.3.2 «Сервисы по управлению электронно-библиотечными системами» прописано, что для уполномоченных представителей образовательной организации должен быть предоставлен доступ с комплексом сервисов управления, таких как отслеживание статистики использования ресурсов, ассоциированный с действующими международными стандартами (по пользователям, изданиям, просмотрам).

Сегодня де-факто стандартом по статистике использования ресурсов является формат COUNTER [**9**], хотя он и не зарегистрирован в международных системах стандартизации в отличие от стандарта SUSHI – протокола передачи статистических данных для автоматического сбора отчётов по ис-

пользованию онлайновых ресурсов, получившего в 2014 г. номер ANSI/NISO Z39.93–2014 [**10**]. Протокол SUSHI описывает автоматическое отправление запроса и получение ответа для статистических отчётов в формате COUNTER.

Согласно определению, представленному в [**11**], COUNTER (https://www.projectcounter.org/) – это стандарт, в котором зафиксированы перечень и форма предоставления статистических данных по использованию электронных ресурсов. Стандарт, известный также как «Свод правил» (*Code of Practice*), позволяет поставщикам и издателям предоставлять библиотекам и провайдерам данных сопоставимые сведения об использовании ресурсов. COUNTER был запущен в марте 2002 г. как международная инициатива по оказанию помощи библиотекарям и издателям в регистрации и обмене статистикой использования электронных ресурсов. По состоянию на июнь 2014 г. сообщество пользователей COUNTER насчитывало около 220 членов и более 50 поставщиков данных, имеющих сертифицированное соответствие одной или нескольким версиям «Свода правил».

В текущей версии 5 Свода правил COUNTER, опубликованных в июле 2017 г. [**9**], оговариваются правила обработки исходных данных (лог-файлов), используемых при составлении отчётов COUNTER.

Среди основных требований отметим следующие:

должны учитываться только успешные и правильные запросы;

для исключения повторного подсчёта следует осуществлять фильтрацию двойного нажатия пользователем той же самой ссылки (такой считается ссылка, если между её повторным нажатием прошло менее 30 секунд).

При составлении отчётов используются идентификаторы единиц контента, называемые элементами (*items*), такие как статьи, главы книг, разделы книг, целые книги, мультимедийный контент. Каждому элементу должен быть присвоен уникальный идентификатор, который привязан к произведению или его части (например, главе или статье) независимо от формата представления (например, PDF, HTML или EPUB). Если в течение одного сеанса пользователь обратился к разным форматам одного и того же элемента, в отчёте должен учитываться только один уникальный вид этого элемента, например PDF-формат.

Федеративный поиск должен учитываться отдельно от поиска, произведённого реальными пользователями, и он фиксируется в отдельном счётчике *Searches_Federated* (для отчётов по использованию баз данных).

Поисковые запросы, осуществлённые через системы Дискавери, равно как и через другие системы, в которых многочисленные базы данных не были выбраны собственно пользователем и которые осуществляют одновременный поиск по нескольким источникам, должны учитываться в отдельном счётчике *Searches_Automated* (для отчётов по использованию баз данных).

Любые запросы, поступающие от интернет-роботов и сканеров, должны быть исключены из отчётов COUNTER.

Отчёты COUNTER не должны включать запросы полнотекстового содержимого, инициированные автоматическими или полуавтоматическими инструментами массовой загрузки, во всех случаях, когда загрузка происходит без прямого вмешательства пользователя.

В результатах отчётов COUNTER используются два варианта работы пользователя с контентом: *исследование* (*Investigation*) и *запрос* (*Request*). В счётчик «*запрос*» включается: просмотр полного текста электронного ресурса в форматах PDF, HTML и пр., а также просмотр присоединённого видео. В счётчик «*исследование*» включаются: счётчик «запрос» + просмотр аннотаций, ссылок на переадресатор протокола OpenURL, просмотр цитируемых источников, ссылок на форму заказа электронной доставки, просмотр предварительной версии статьи. В отчётах фиксируются следующие суммарные значения:

суммарное число запрошенных единиц контента в варианте «исследование» (*Total_Item_Investigations*);

суммарное число переданных единиц контента в формате «запрос» (*Total_Item_Requests*);

число уникальных единиц контента, запрошенных в варианте «исследование» (*Unique_Item_Investigations*);

число уникальных единиц контента, переданных в варианте «запрос» (*Unique_Item_Requests*);

число уникальных заглавий, запрошенных в варианте «исследование» (*Unique_Title_Investigations*);

число уникальных заглавий, переданных в варианте «запрос» (*Unique_Title_Requests*).

Отчёты об использовании содержимого журнала формируются без учёта статей, предоставляемых в варианте *Gold Open Access,* которые учитываются отдельно. «Золотой» открытый доступ означает, что журнал не требует денег за доступ читателя к опубликованной в нём электронной статье. Проблема «золотого» доступа состоит в том, что издание требует значительной оплаты от автора публикации. По сути это чисто коммерческая модель, ещё более выгодная для издателей, чем модель подписки [**12**].

Особенность отчётов COUNTER, как отмечено в [**7**], – отсутствие понятия *пользователь*. Посещения COUNTER учитывает только в формате поисковых запросов *Total_Item_Investigations*, *Total_Item_Requests.* Однако следует отметить, что практически все поставщики контента, поддерживающие этот стандарт, также ведут статистику количеству и пользователей, и посещений (сессий) в дополнительных отчётных формах. Как отмечалось в

[**13**], важно адаптировать международную форму библиотечной статистики в части электронных ресурсов (COUNTER) с учётом действующего российского законодательства и специфических российских форм функционирования электронных ресурсов. В [**14**] приведены конкретные предложения по дополнению отчётов COUNTER для электронных книг и статей.

Для получения объективной статистики по обращению к электронным каталогам и к отдельным электронным ресурсам следует, по нашему мнению, дополнить приведённый выше список следующими показателями:

число запросов пользователей к отдельной единице контента, поступивших от организации (определяется по IP-адресу организации);

число запросов к отдельной единице контента, поступивших от интернет-роботов и сканеров;

число запросов к отдельной единице контента, поступивших от программных продуктов, осуществляющих федеративный поиск.

Подводя итоги, следует отметить, что в настоящее время в библиотеках России нет единой методики подсчёта обращений к электронным ресурсам, которая, с одной стороны, соответствовала бы требованиям ГОСТов, а с другой – поддерживала совместимость с международными стандартами, такими как COUNTER. В то же время разработчики систем автоматизации библиотек имеют все возможности для постепенного практического внедрения таких расчётов в свои программные продукты.

Разумное сочетание требований ГОСТов, базовых показателей COUNTER, а также расчёты дополнительных показателей, полученных на основе анализа лог-файлов веб-серверов библиотек, дают богатые возможности для исследования востребованности электронных ресурсов библиотек различными категориями пользователей и являются ещё одним значимым источником данных при расчёте альтметрик.

СПИСОК ИСТОЧНИКОВ

1. **Гончаров М. В.** Электронная библиотека ГПНТБ России: динамика пополнения, технологии, ресурсы / М. В. Гончаров, К. А. Колосов // Науч. и техн. б-ки. – 2018. – № 12. – С. 34–41.

Goncharov M. V. *Elektronnaya biblioteka GPNTB Rossii: dinamika popolneniya, tehnologii, resursy / M. V. Goncharov, K. A. Kolosov // Nauch. i tehn. b-ki. – 2018. – № 12. – S. 34–41.*

2. **Гончаров М. В., Михайленко И. И.** Интеграция информационных ресурсов ГПНТБ России в рамках Системы открытого архива // Там же. – № 4. – С. 5–13.

Goncharov M. V., Mihaylenko I. I. *Integratsiya informatsionnyh resursov GPNTB Rossii v ramkah Sistemy otkrytogo arhiva // Tam zhe. – № 4. – S. 5–13.*

3. **Гончаров М. В., Колосов К. А.** Разработка системы открытого архива ГПНТБ России // Там же. – № 12. – С. 42–48.

Goncharov M. V., Kolosov K. A. *Razrabotka sistemy otkrytogo arhiva GPNTB Rossii // Tam zhe. – № 12. – S. 42–48.*

4. **Земсков А. И.** Библиометрия в библиотеках / А. И. Земсков, К. А. Колосов // Там же. – 2016. – № 11. – С. 5–23.

Zemskov A. I. *Bibliometriya v bibliotekah / A. I. Zemskov, K. A. Kolosov // Tam zhe. – 2016. – № 11. – S. 5–23.*

5. **Юркевич М. А.** Перспективы применения альтметрики в социогуманитарных науках / М. А. Юркевич, И. П. Цапенко // Информ. о-во. – 2015. – № 4. – С. 9–16.

Yurkevich M. A. *Perspektivy primeneniya altmetriki v sotsiogumanitarnyh naukah / M. A. Yurkevich, I. P. Tsapenko // Inform. o-vo. – 2015. – № 4. – S. 9–16.*

6. **ГОСТ Р 7.0.20–2014** Библиотечная статистика: показатели и единицы исчисления [Электронный ресурс]. – Режим доступа: http://docs.cntd.ru/document/1200113790.

GOST R 7.0.20–2014 *Bibliotechnaya statistika: pokazateli i edinitsy ischisleniya [Elektronnyy resurs].*

7. **Белов А. М.** Библиотечная статистика сетевых ресурсов≠ Статистика сетевых ресурсов в библиотеке? / А. М. Белов // Б-ки вузов Урала. – 2015. – № 14. – С. 123–129.

Belov A. M. *Bibliotechnaya statistika setevyh resursov ≠ Statistika setevyh resursov v biblioteke? / A. M. Belov // B-ki vuzov Urala. – 2015. – № 14. – S. 123–129.*

8. **ГОСТ Р 57723-2017** Информационно-коммуникационные технологии в образовании. Системы электронно-библиотечные. Общие положения. 2017 [Электронный ресурс]. – Режим доступа: http://docs.cntd.ru/document/1200156825.

GOST R 57723-2017 *Informatsionno-kommunikatsionnye tehnologii v obrazovanii. Sistemy elektronno-bibliotechnye. Obshchie polozheniya. 2017 [Elektronnyy resurs].*

9. **COUNTER** Code of Practice: Release 5 [Электронный ресурс]. – URL: https://www.projectcounter.org/wp-content/uploads/2017/10/ Release5_20171013-1.pdf.

10. **Standardized** Usage Statistics Harvesting Initiative (SUSHI) Protocol (ANSI/NISO Z39.93-2014) [Электронный ресурс]. – URL: https://www.niso.org/standards-committees/sushi.

11. **Методические** рекомендации по разработке репозиториев / под ред. М. Е. Шварцмана. – Москва : Ваше цифровое изд-во, 2018. – 34 с. – ISBN 978-5-6040408-2-9.

Metodicheskie *rekomendatsii po razrabotke repozitoriev / pod red. M. E. Shvartsmana. – Moskva : Vashe tsifrovoe izd-vo, 2018. – 34 s. – ISBN 978-5-6040408-2-9.*

12. **Шрайберг Я. Л.** Модели открытого доступа: история, виды, особенности, терминология / Я. Л. Шрайберг, А. И. Земсков // Науч. и техн. б-ки. – 2008. – № 5. – С. 68–79.

Shrayberg Ya. L. *Modeli otkrytogo dostupa: istoriya, vidy, osobennosti, terminologiya / Ya. L. Shrayberg, A. I. Zemskov // Nauch. i tehn. b-ki. – 2008. – № 5. – S. 68–79.*

13. **Билан И. В.** Статистика использования электронных ресурсов в библиотеке [Электронный ресурс]. – Режим доступа: http://www.gpntb.ru/libcom11/disk/14.pdf.
Bilan I. V. *Statistika ispolzovaniya elektronnyh resursov v biblioteke [Elektronnyy resurs].*

14. **Давыдова Н. Р.** Что считать при использовании электронных ресурсов? [Электронный ресурс]. – Режим доступа: https://textualheritage.org/ru/el-manusctipt-08-/71.html.
Davydova N. R. *Chto schitat pri ispolzovanii elektronnyh resursov? [Elektronnyy resurs].*

Mikhail Goncharov*, Cand. Sc. (Technology), Associate Professor, Leading Researcher; Head, Perspective Research and Analytic Forecasting Group, Russian National Public Library for Science and Technology;*
goncharov@gpntb.ru
17, 3rd Khoroshevskaya st., 123298 Moscow, Russia

Kirill Kolosov*, Cand. Sc. (Technology), Leading Researcher, Russian National Public Library for Science and Technology;*
kolosov@gpntb.ru
17, 3rd Khoroshevskaya st., 123298 Moscow, Russia