UKSG

Open access book usage data – how close is COUNTER to the other kind?

In April 2020, the OAPEN Library moved to a new platform, based on DSpace 6. During the same period, IRUS-UK started working on the deployment of Release 5 of the COUNTER Code of Practice (R5). This is, therefore, a good moment to compare two widely used usage metrics – R5 and Google Analytics (GA). This article discusses the download data of close to 11,000 books and chapters from the OAPEN Library, from the period 15 April 2020 to 31 July 2020. When a book or chapter is downloaded, it is logged by GA and at the same time a signal is sent to IRUS-UK. This results in two datasets: the monthly downloads measured in GA and the usage reported by R5, also clustered by month. The number of downloads reported by GA is considerably larger than R5. The total number of downloads in GA for the period is over 3.6 million. In contrast, the amount reported by R5 is 1.5 million, around 400,000 downloads per month. Contrasting R5 and GA data on a country-by-country basis shows significant differences. GA lists more than five times the number of downloads for several countries, although the totals for other countries are about the same. When looking at individual tiles, of the 500 highest ranked titles in GA that are also part of the 1,000 highest ranked titles in R5, only 6% of the titles are relatively close together. The choice of metric service has considerable consequences on what is reported. Thus, drawing conclusions about the results should be done with care. One metric is not better than the other, but we should be open about the choices made. After all, open access book metrics are complicated, and we can only benefit from clarity.

Keywords

open access books; usage data; COUNTER; Google analytics; downloads; impact measurement

Introduction

Since its launch in 2010, the OAPEN Library has made peer-reviewed books and chapters available in open access (OA).[1] By February 2021, the collection had grown to 14,500 books and over 700 chapters. Starting in June 2013, IRUS-UK provided us with COUNTER Release 4 compliant usage data for the OAPEN Library.[2] The Library passed the ten million downloads mark in the first quarter of 2020.

In April 2020, the OAPEN Library moved to a new platform, based on DSpace 6, the open source repository system. Among other things, this allowed us to monitor all events happening on the platform using Google Analytics (GA). During the same period, IRUS-UK started working on the deployment of Release 5 (R5) of the COUNTER Code of Practice.

This is, therefore, a good moment to compare these two widely used usage metrics. By describing the OAPEN Library usage data from Google Analytics and COUNTER Release 5 we aim to better understand the differences. We do not mean to make judgement as we do not think one is better than the other. These systems are developed from a different perspective: while GA is optimized to describe what is happening on a certain website – especially from a marketing and sales perspective – COUNTER aims to provide standardized data that can be used to aggregate and compare across multiple environments.

RONALD SNIJDER

Deputy Director
OAPEN Foundation

'By describing the OAPEN Library usage data from Google Analytics and COUNTER Release 5'

Literature review

This is far from the first article examining the usage data of open access books. The usage data can be seen as an indicator of their impact: the geographical spread and the number of downloads are often used as indicators. Apart from downloads, citation data and altmetrics are also of interest to researchers and several publishers have investigated the impact of open access on their books.

In a case study of UCL Press, Montgomery et al.[3] compared several sources of download figures to understand how they are affected by significant events related to the promotion of published titles. GA was not set up to record download figures but was used here to provide information about the visitors to the UCL website. The study placed much emphasis on the fact that each platform provides usage data based on different principles and the download figures from the three repositories were therefore not aggregated.

Stockholm University Press analysed usage statistics, citation data and altmetrics, in combination with a survey of attitudes and behaviour among authors and editors who have published open access books.[4] The authors came to the conclusion that there are differences within specific academic disciplines but also mentioned that interpretation of the metrics is still complicated.

Springer Nature undertook a case study – based on 3,934 books, including 281 OA books – examining the differences in impact of books published in open access compared to books that were published in a closed manner.[5] The authors concluded that making books open access increased the number of downloads, and also that the geographical spread – especially downloads from low- and middle-income countries – also expanded. Furthermore, open access books were also cited twice as much compared to their 'closed' counterparts.

Recently, Taylor[6] researched the number of times open access books are mentioned in social networks, mass media and blogs and in policy documents. According to the author, there is an 'open access advantage', but at this moment, the underlying mechanisms are not clear. Again, differences between academic disciplines are visible.

Another attempt at understanding the impact of open access books is the article by Snijder.[7] By categorizing the users, the author aimed to gather quantitative data about the scientific impact and societal relevance of the downloaded titles. From the measured data, over 27% was directly linked to academic users while more than 45% of the downloads have a high probability of coming from the general public or other non-academics – a possible indication of societal impact.

Ozaygen[8] has written an extensive technical analysis of open access usage data of a collection of 28 newly published open access books in several academic disciplines, provided by 13 publishers. This was the pilot collection of the Knowledge Unlatched[9] initiative, made available in 2013. It combines several techniques to provide a comprehensive picture of how – and where – the books were used or made available on the web.

'Open Access eBook Usage (OAeBU) Data Trust for usage data on open access books'

In order to find and analyse the impact of a particular open access book, one needs to spend quite a lot of time and effort. To help solve this problem, the Open Access eBook Usage (OAeBU) Data Trust[10] is being established. It is a two-year pilot to develop and test infrastructure, policy and governance models to create a global data trust for usage data on open access books.[11] Apart from collecting usage data, the data trust aims to align with the priorities of authors and institutions while respecting ethical norms in the use of metrics.

The COUNTER code of practice is intended to provide libraries with consistent and comparable statistics about the online resources they procure.[12] Libraries not only need to measure and evaluate how their external resources are used, as there might be resources whose prices depend on the use, but also to quantify the role of the library itself.[13] Libraries are also using GA as a tool to help visualize how their website – including the library catalogue – is used.[14]

From this literature review, we can conclude that the usage data of open access books and chapters plays an important role for both publishers and libraries. It is also clear that obtaining the data is far from easy, and – on top of that – there are still a lot of uncertainties: differences in types of data (downloads, citations, altmetrics) coming from multiple platforms that might generate incomparable data. Added to that, is the necessity to interpret the outcomes. The next section will illustrate the differences between two widely used metrics, reporting on the same event: downloading books and chapters from the OAPEN Library.

'usage data of open access books and chapters plays an important role for both publishers and libraries'

The data

This article discusses the download data of close to 11,000 books and chapters from the OAPEN Library, from the period 15 April 2020 to 31 July 2020. When a book or chapter is downloaded, it is logged by GA and at the same time a signal is sent to IRUS-UK. The reported results have been used for the comparison in this article.

GA logs many more things than downloads: it captures all visits to a website and collects information about the visitors. The challenge is to only find the usage data that is relevant for this comparison. We created a customized report that captures downloads – not web page visits. In GA, this is termed an 'event' in the category 'Bitstream'. The OAPEN DSpace environment does not only contain 'book files' but each title is also accompanied by a cover image file. We excluded the downloads of cover files from the reports. Furthermore, to ensure that comparable data is used, known 'bots' are filtered out.

The data gathering of IRUS-UK is purely focused on usage of publications. The downloads are assessed according to the R5 guidelines and are reported as an 'Item Filter Report'. Here, we used the metric 'Total_Item_Request', which is defined as the total number of times the full text of a content item was downloaded or viewed. Crucial to COUNTER reporting is the removal of any usage data that is deemed to be unintended by a – human – user.[15] Thus, automated downloads by 'bots' is excluded.

Both the GA and the R5 platforms offer the possibility to deduplicate usage data, called 'Unique Events' in GA and 'Unique_Item_Requests' in R5. As we could not be certain that both platforms use the same definitions of a unique event, we decided not to use this metric. The selection choices are listed in Table 1.

	Google Analytics	**COUNTER R5**
Supplied by	Google	IRUS-UK, Jisc
Report/filter	• Event category: Bitstream	• COUNTER R5 Item Report
	• Filters used:	• Filters used:
	◦ Exclude '.jpg' from Page (part of the URL after the domain name)	◦ Repository: OAPEN DSpace
	◦ Bot filtering: exclude all hits from known bots and spiders	
Data used	Total Events	Total_Item_Requests
Frequency	Daily	Monthly
Period	15 April 2020 to 31 July 2020	15 April 2020 to 31 July 2020
Description	[16]	[17]

Table 1. GA and R5 data selection

This results in two datasets: the monthly downloads measured in GA and the usage reported by R5, also clustered by month. Both datasets consist of the total number of downloads per title, broken down per country. So, in July 2020, according to the R5 data, the book *Ethnicity, Race and Inequality in the UK*[18] was downloaded 1,433 times, and the readers resided in 54 different countries. When we look at the GA data, the picture is a little different: 1,360 downloads coming from 21 countries.

In this example, the difference between R5 and GA is relatively small, but there is usually a significant discrepancy between the two datasets. In general, to be COUNTER compliant, usage data must conform to stricter rules to be reported when compared to the GA measurements. When the total number of downloads is compared, the R5 data is 58% of the GA total.

In the following sections, we will compare the GA and the R5 data on several levels: starting from the totals, via the country data to a comparison at book level. All data are available – details may be found in the data accessibility statement at the end of this article.

Total usage

As mentioned before, the number of downloads reported by GA is considerably larger than R5. The total number of downloads in GA is over 3.6 million: more than 1 million downloads per month. In contrast, the amount reported by R5 is 1.5 million downloads: around 400,000 downloads per month.

'the relation between GA and R5 is not completely straightforward'

When looking at the monthly data – as depicted in Figure 1 – it becomes clear that the relation between GA and R5 is not completely straightforward. The percentage difference varies from month to month: in May the difference was 54%, and this climbed to 64% in July. Of course, three months is not enough to declare a trend, but it would be interesting to conduct another analysis after a year.
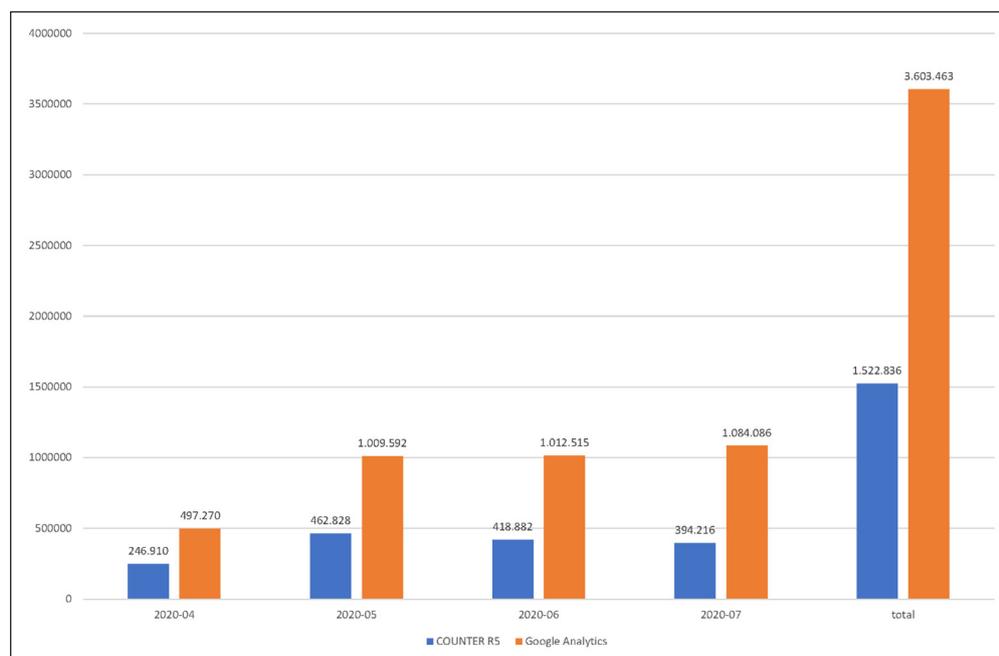

Figure 1. Comparing R5 to GA

Country comparison

If the total usage data were to be broken down by country and projected on a map of the world, it would be difficult to see significant differences: both would display usage in almost every country. Both GA and R5 list downloads from Afghanistan to Zimbabwe. Also, the data follow the same pattern: a few countries where relatively many books and chapters are downloaded, and a 'long tail' of countries.

'Open access is clearly a global phenomenon, not limited to the most affluent countries'

It is more interesting to look at the differences between the 'top 15': United States, Germany, United Kingdom, France, India, Australia, China, the Netherlands, Russia, Indonesia, Canada, Italy, South Africa, Austria and Spain. Open access is clearly a global phenomenon, not limited to the most affluent countries. Comparing the total number of downloads for these countries leads to a familiar conclusion: the R5 total is 58% of the GA total. This is in line with the pattern for total usage.

However – as is shown in Figure 2 – contrasting R5 and GA data on a country-by-country level shows significant differences. GA lists more than five times the number of downloads for the USA, France, China and Russia. In contrast, the numbers for Australia, Canada and Austria are about the same.
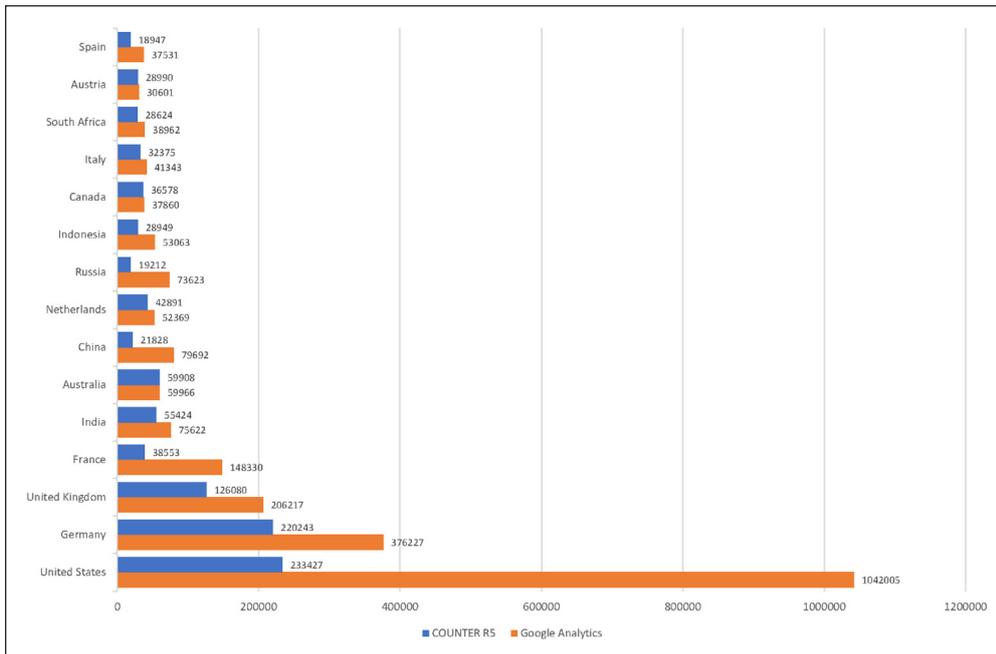

Figure 2. Comparing usage of the top 15 countries

Figure 3 shows the differences between GA and R5 in a slightly different way. According to the GA data, usage is dominated by U.S. based addresses. Here, the American downloads are almost a third of the total, three times as many as Germany, the second country. The R5 data paints a more 'balanced' picture, where the differences between the countries with the most downloads are much smaller.
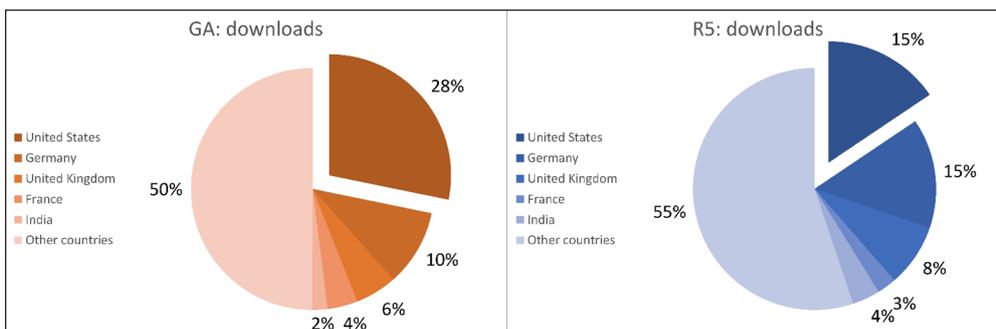
Figure 3. Percentage of top five countries in GA and R5

Differences at title level

The last level to be discussed is the differences between GA and R5 when looking at individual titles. Given the fact that our datasets contain nearly 11,000 titles, a thorough discussion of each title would be repetitive and not very helpful. However, comparing the ranking of the titles helps to create a picture. Simply put: each book is ranked according to the number of downloads, where the book with the highest number of downloads is ranked at number one, and so forth. The next step is to compare the ranking of the titles: the differences of the ranking by GA and R5 indicate how the usage of the book or chapter is depicted.

'comparing the ranking of the titles helps to create a picture'

A first indication of the large discrepancies between GA and R5 is illustrated by Table 2. When looking at the 500 highest ranked titles in GA that are also part of the 1,000 highest ranked titles in R5, only 6% of the titles are relatively close together. An example of this

would be the book *Frankenstein*,[19] ranked fifth in GA and ranked third in R5. Also, a relatively small part – 20% – consists of books and chapters that are ranked within 50 places of each other. Here, the book *The Myths That Made America*[20] can be used as an illustration: ranked sixth in GA and ranked twenty-third in R5. The largest group of titles is ranked further apart, such as *Health of People, Health of Planet and Our Responsibility*[21] which is ranked second in GA but ranked eighty-third in R5.

Ranking	Number of titles	Percentage
Difference < 10	30	6%
Difference < 50	98	20%
Larger difference	372	74%
Total	**500**	**100%**

Table 2. Ranking of GA and R5 compared

The differences in ranking are even more striking when they are visualized. In Figure 4, 50 books are represented as coloured bars. The length of the bar corresponds with the rank. Many titles with a high rank in GA are lowly ranked in R5 and vice versa, without any apparent underlying pattern.
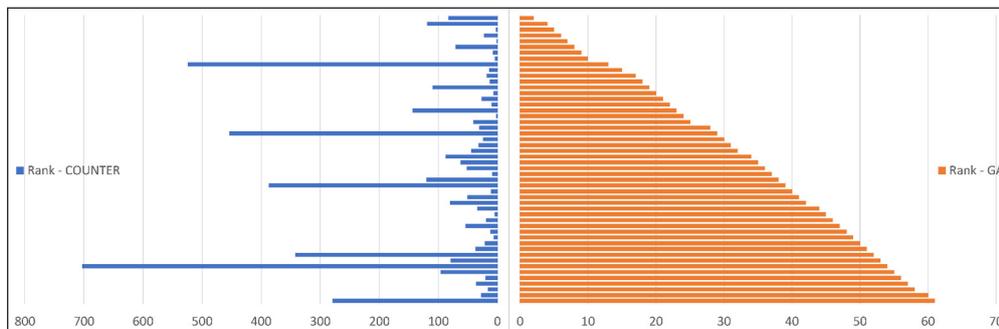

Figure 4. Ranking of 50 titles in GA and R5

The following subsections describe the five highest ranked books in GA with their R5 'counterpart data'. We will see that each title's usage is represented quite differently in GA and R5 and that the number one ranked title in GA – *m-Learning – die neue Welle? Mobiles Lernen für Deutsch als Fremdsprache*[22] – was not found in the first 1000 ranked titles in R5. Therefore, the first title in the next section is the second-most downloaded title in the GA data.

Health of People, Health of Planet and Our Responsibility

For this title, GA reports close to 28,000 downloads. In contrast, R5 only reports just over 1,200 downloads, which leads to a large disparity in ranking. See Table 3 and Figure 5.

	GA	R5
Rank	2	83
Number of downloads	27,962	1,274

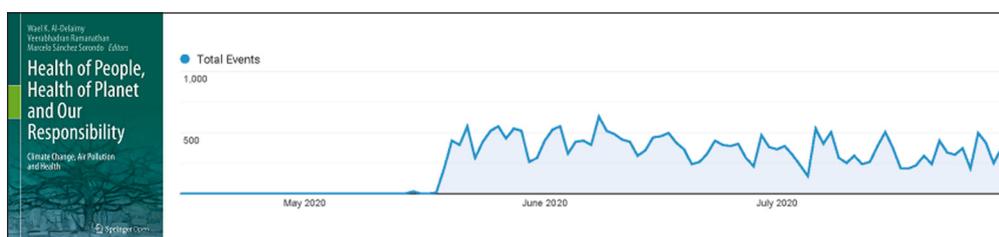Table 3. Ranking and downloads of *Health of People, Health of Planet and Our Responsibility*

Figure 5. GA usage of *Health of People, Health of Planet and Our Responsibility*

Access Controlled

In the case of this book, over 7,000 of the downloads took place on one day. IRUS-UK filters out users who download 40 or more publications in a single day or those that download the same publication more than 10 times in a single day. See Table 4 and Figure 6.

	GA	R5
Rank	4	119
Number of downloads	12,457	1,083

Table 4. Ranking and downloads of *Access Controlled*
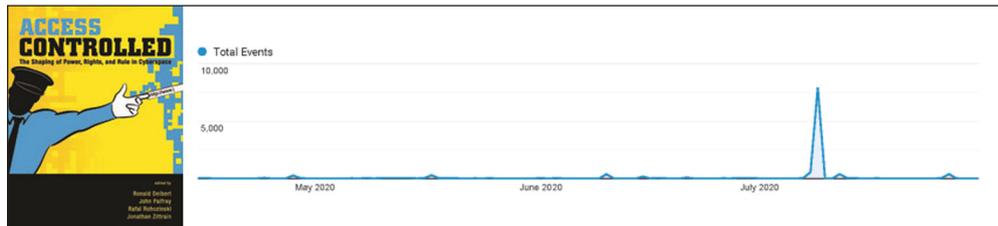
Figure 6. GA usage *of Access Controlled*

Frankenstein

Both GA and R5 report many downloads, leading to a high ranking on both 'sides'. The download pattern is in stark contrast with the 'single day download' of *Access Controlled*. See Table 5 and Figure 7.

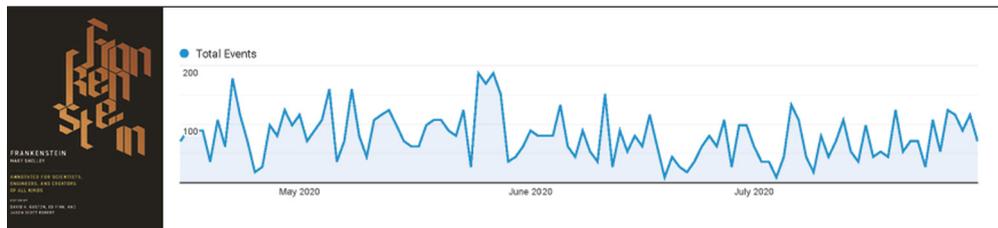|	GA	R5
Rank	5	3
Number of downloads	8,583	5,675

Table 5. Ranking and downloads of *Frankenstein*

Figure 7. GA usage of *Frankenstein*

The Myths That Made America

At first glance, the usage pattern looks a lot like *Access Controlled*: peak usage in a short period. However, the peaks are not as large: the highest number of downloads in one day did not exceed 1,300, a lot less than the 7,000 downloads of *Access Controlled*. See Table 6 and Figure 8.

	GA	R5
Rank	6	23
Number of downloads	8,299	2,098

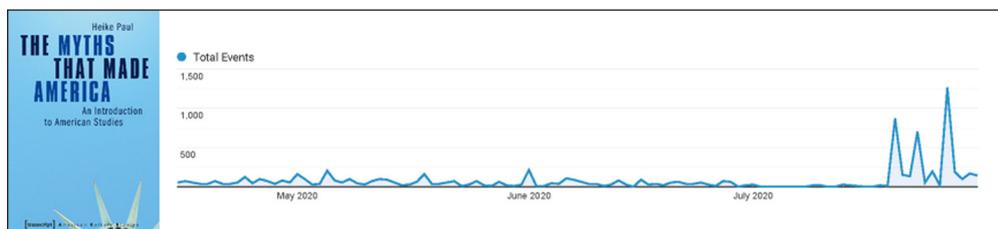Table 6. Ranking and downloads of *The Myths That Made America*

Figure 8. GA usage of *The Myths That Made America*

Ethnicity, Race and Inequality in the UK

For this title, the number of downloads recorded by GA is actually lower than those recorded by R5. This is an interesting reversal of the pattern. See Table 7 and Figure 9.

	GA	R5
Rank	7	1
Number of downloads	7,828	10,511

Table 7. Ranking and downloads of *Ethnicity, Race and Inequality in the UK*
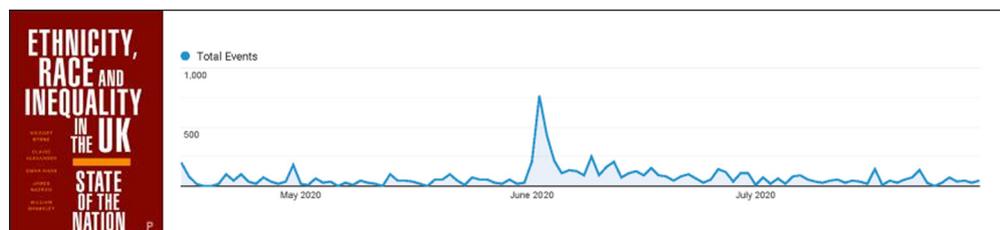
Figure 9. GA usage of *Ethnicity, Race and Inequality in the UK*

Conclusions

Usage data of open access books are important to many stakeholders. However, there is no universally accepted standard that is used by all providers of OA book collections. Apart from differences in the data provided, collecting the data is not an easy task.

This article aims to display how two widely used metrics services – Google Analytics and COUNTER Release 5 – report about the same events. Both services have made their own choices on what is reported, and what is not. There were significant discrepancies seen during the period studied: GA reported 3.6 million downloads in contrast to the 1.6 million downloads stated by R5. Moreover, there is no simple rule of thumb to 'convert' GA metrics to R5: at the level of country totals and at the level of the individual titles, we can see wildly different figures. For instance, the usage data as reported by GA compared to R5 is much higher for the USA, while the data for Australia is virtually the same. This also holds true for the book *Access Controlled* versus *Frankenstein*.

It may be tempting to conclude that the usage as reported by GA is 'truer', as it seems to have fewer restrictions on what is measured. That is not the case. First of all, the GA data used in this comparison already used a filter to remove usage from known 'bots'. Secondly, as the example of the book *Ethnicity, Race and Inequality in the UK* has shown, the usage reported by GA may sometimes be more constrained than R5.

'One metric is not better than the other, but we should be open about the choices made'

What became very clear is that the choice of metric service has considerable consequences for what is reported. Thus, drawing conclusions about the results should be done with care. For instance, what should be made of the fact that the most downloaded title according to GA was not even found in the 1,000 most downloaded titles according to R5?

One metric is not better than the other, but we should be open about the choices made. After all, open access book metrics are complicated and we can only benefit from clarity.

References

1. "OAPEN Library", *OAPEN Foundation*, 2010,
    https://www.oapen.org (accessed 15 February 2021).

2. "International Collaboration and Value: Working with OAPEN", *IRUS-UK Case Studies*, 2018,
    https://irus.jisc.ac.uk/documents/IRUS-UK_working_with_OAPEN.pdf (accessed 15 February 2021).

3. Lucy Montgomery et al., "Getting the Best out of Data for Open Access Monograph Presses: A Case Study of UCL Press," *Learned Publishing* 31, no. 4 (1 October 2018): 335–44, DOI:
    https://doi.org/10.1002/leap.1168 (accessed 17 February 2021).

4. Sofie Wennström et al., "The Significant Difference in Impact: An Exploratory Study about the Meaning and Value of Metrics for Open Access Monographs," *ELPUB 2019 23d International Conference on Electronic Publishing* (June 2019), DOI:
    https://doi.org/10.4000/proceedings.elpub.2019.9 (accessed 15 February 2021).

5. Ros Pyne et al., "The Future of Open Access Books: Findings from a Global Survey of Academic Book Authors," *Springer Nature Open Access Books*, (June 2019), DOI:
    https://doi.org/10.6084/m9.figshare.8166599 (accessed 15 February 2021).

6. Megan Taylor, "Mapping the Publishing Challenges for an Open Access University Press," *Publications* 7, no. 4 (December 2019): 63, DOI:
    https://doi.org/10.3390/publications7040063 (accessed 17 February 2021).

7. Ronald Snijder, "Measuring Monographs: A Quantitative Method to Assess Scientific Impact and Societal Relevance," *First Monday* 18, no. 5 (May 2013), DOI:
    https://doi.org/10.5210/fm.v18i5.4250 (accessed 17 February 2021).

8. Alkim Ozaygen, "Analysing the Usage Data of Open Access Scholarly Books: What Can Data Tell Us?," (Thesis, Curtin University, 2019),
    https://espace.curtin.edu.au/handle/20.500.11937/79585 (accessed 17 February 2021).

9. Lucy Montgomery and Celeste Feather, "Knowledge Unlatched Pilot Collection to Become Open Access – Nearly 300 Libraries Globally Pledge Their Support," *Knowledge Unlatched* (blog), 10 March 2014,
    https://knowledgeunlatched.org/2014/03/knowledge-unlatched-pilot-collection-to-become-open-access-nearly-300-libraries-globally-pledge-their-support/ (accessed 19 February 2021).

10. "Developing a Pilot Data Trust for Open Access Ebook Usage," Educopia Institute,
    https://educopia.org/data_trust/ (accessed 17 February 2021).

11. Christina Drummond, "Engaging Stakeholder Networks to Support Global OA Monograph Usage Analytics," *Collaborative Librarianship* 12, no. 2 (21 October 2020),
    https://digitalcommons.du.edu/collaborativelibrarianship/vol12/iss2/9 (accessed 15 February 2021); "Developing a Pilot Data Trust for Open Access Ebook Usage," Educopia Institute,
    https://educopia.org/data_trust/ (accessed 15 February 2021).

12. "COUNTER │ About Us", *COUNTER*, 2014,
    http://www.projectcounter.org/about.html (accessed 15 February 2021).

13. Gayle Baker and Eleanor J. Read, "Vendor-Supplied Usage Data for Electronic Resources: A Survey of Academic Libraries," *Learned Publishing* 21, no. 1 (1 January 2008): 48–57, DOI:
    https://doi.org/10.1087/095315108X247276 (accessed 15 February 2021); Virginia Kinman, "E-Metrics and Library Assessment in Action," *Journal of Electronic Resources Librarianship* 21, no. 1 (16 April 2009): 15–36, DOI:
    https://doi.org/10.1080/19411260902858318 (accessed 15 February 2021).

14. Tabatha Farney and Nina McHale, "Introducing Google Analytics for Libraries," *Library Technology Reports* 49, no. 4 (2013): 5–8,
    https://www.journals.ala.org/index.php/ltr/article/view/4269 (accessed 15 February 2021); Steven J. Turner, "Website Statistics 2.0: Using Google Analytics to Measure Library Website Effectiveness," *Technical Services Quarterly* 27, no. 3 (May 2010): 261–78, DOI:
    https://doi.org/10.1080/07317131003765910 (accessed 15 February 2021).

15. "7.0 Processing Rules for Underlying COUNTER Reporting Data," *COUNTER*,
    https://www.projectcounter.org/code-of-practice-five-sections/7-processing-rules-underlying-counter-reporting-data/ (accessed 15 February 2021).

16. 'About Events – Analytics Help', *Google*, 2020,
    https://support.google.com/analytics/answer/1033068?hl=en#Anatomy (accessed 15 February 2021).

17. "Release 5: Understanding Investigations and Requests," *COUNTER*, 26 May 2017,
    https://www.projectcounter.org/release-5-understanding-investigations-and-requests/ (accessed 15 February 2021).

18. Bridget Byrne et al., *Ethnicity, Race and Inequality in the UK: State of the Nation* (Bristol: Policy Press, 2020),
    https://library.oapen.org/handle/20.500.12657/22310 (accessed 15 February 2021).

19. Ed Finn, David Guston, and Jason Scott Robert, *Frankenstein : A New Edition for Scientists and Engineers* (Cambridge, Massachusetts: The MIT Press, 2017),
    https://library.oapen.org/handle/20.500.12657/31387 (accessed 15 February 2021).

20. Heike Paul, *The Myths That Made America : An Introduction to American Studies* (transcript Verlag, 2014),
    https://library.oapen.org/handle/20.500.12657/31456 (accessed 15 February 2021).

21. Wael Al-Delaimy, Veerabhadran Ramanathan, and Marcelo Sánchez Sorondo, *Health of People, Health of Planet and Our Responsibility: Climate Change, Air Pollution and Health* (Cham: Springer Nature, 2020),
    https://library.oapen.org/handle/20.500.12657/37742, DOI:
    https://doi.org/10.1007/978-3-030-31125-4 (accessed 15 February 2021).

22. Haymo Mitschian, *m-Learning – die neue Welle? Mobiles Lernen für Deutsch als Fremdsprache* (Kassel: Kassel University Press, 2010), DOI:
    https://doi.org/10.26530/OAPEN_355293 (accessed 15 February 2021).

Ronald Snijder
Deputy Director
OAPEN Foundation, NL
E-mail: r.snijder@oapen.org
ORCID ID: 0000-0001-9260-4941