# RDA Data Usage Metrics WG Recommendations

Winter, 2021

**Background & Summary**

Research data are increasingly recognized as important outputs of scholarly research, yet there are currently no standardized or comprehensive metrics for research data as there are for articles. This Working Group was founded following a Birds of a Feather at RDA Plenary 10 hosted by the Make Data Count initiative. Lending expertise from various projects and research stakeholders, this WG, a part of the Publishing Data IG, aimed to harness community buy-in of standardized approaches to data usage metrics and drive widespread adoption. The first WG meeting at RDA11 focused on an overview of initiatives in the data usage metrics space and spent the majority of the time discussing scope for the WG. Two virtual meetings took place before RDA12 focused on refining scope and defining data usage metrics. The RDA12 session centered on use cases for usage metrics, updates to the COUNTER Code of Practice for Research Data, and a discussion on barriers to adoption of standardized usage metrics. RDA13 had the largest attendance yet, overflowing a room as we presented on survey results of current implementations of usage metrics and barriers to adoption. At RDA14, WG members presented on the pitfalls and shortcomings of data usage metrics and further analyses of the survey. A discussion began about where the WG should head, sharing and developing practices around data usage metrics. The last WG session at RDA16 gave an opportunity for a split crowd of new and returning members to give input on the proposed recommendations below. The broad takeaway is that community-agreed usage metrics are essential for the future of research data evaluation, but technical, bibliometric, and social infrastructure are required to properly develop indicators.

**Outputs**

- Data Usage Metrics at Repositories: A Survey -- https://zenodo.org/record/3476545
- Potential Shortcomings Review (video) -- https://vimeo.com/367997861
- Potential Shortcomings Review (document) -- https://gitlab.lrz.de/RDA/data-usage-metrics-wg/-/blob/master/shortcomings.md

**Recommendations**

*Premise 1: Data usage is a topic the community is broadly interested in, but data usage is complicated and there is little activation energy to tackle some of the complex issues. While these complex areas will need to be resolved, focus should remain on adoption of minimal frameworks for counting data usage.*
The WG had broad attendance at plenaries but had little support in virtual meetings between plenaries. This suggests that while folks want to hear and give input on use cases for data usage metrics, few are interested in developing solutions. The group had intended to develop a series of use cases followed by identified types of usage metrics for each case. This quickly evolved as we did not have WG participation to develop these use cases broadly. After collecting use cases from some, we also recognized that this was not a comprehensive approach as we did not have

bibliometrics participation in the WG to properly research and assess types of metrics that may be a good fit for the various use cases. The survey we released, which garnered much attention, reflects hesitancy to implement first steps towards metrics (e.g., normalization against a standard) due to resourcing. Complex issues have been noted many times like dataset granularity and dataset versioning but the group agrees that basic adoption of practices is the most important thing at this point. Going forward, complex and trickier issues should be documented and discussed in a non-distracting way from the first goal of broad implementation of normalized data usage counting. Sub-communities with common concerns or needs that are not accommodated by the basic use cases should be identified across and within disciplines, and should work together to build from these general use cases, and develop additional best practices that accommodate their specific requirements.

*Premise 2: The RDA Data Usage Metrics WG supports the Make Data Count recommendations for normalized data usage, utilizing the COUNTER Code of Practice for Research Data Code of Practice, and utilizing DataCite for aggregations*

Repositories should use the COUNTER Code of Practice as a starting point for normalizing the counting process for data usage. When issues arise, or use cases that do not fit the CoP are found, these should be collected through the WG and through Make Data Count channels. This will allow for the Code of Practice to incorporate feedback and new details in future iterations. This first step of ensuring repositories are counting in the same way will be a step forward in making dataset usage comparable across repositories. For these comparisons and bibliometrics studies that are required for the development of data metrics to occur, repositories that are clients of DataCite should submit their usage to the EventData aggregate hub.

*Premise 3: The RDA and larger research data communities should be cautioned against defaulting to data metrics like a data impact factor*

Considering the lack of developments in this area beyond the Make Data Count initiative, this WG feels strongly that defaulting to established article level metrics for data will be harmful for the research data landscape. Instead, interested communities should focus on agreed upon building blocks for assessing and comparing data usage, for instance, normalizing how data usage is counted and exposed. Jumping to indicators like a data impact factor or data h index would not drive this community forward as there are not enough bibliometric studies available yet to understand what the right indicators for various types of research data may be.

*Premise 4: The WG and larger community should consider potential shortcomings, mitigations that can be made, and broader use cases for implementation to better understand usage types*

Understanding that the main goal should be adoption of comparable data usage counts, there are important considerations to investigate like:

- Datasets that do not use DOIs and instead use other IDs such as accession numbers
- Reuse of datasets that does not involve a download that can be readily counted
- Datasets that are limited in their public access (e.g., behind institutional firewalls)

- Accounting for potential shortcomings through bibliometrics studies and further understanding of how datasets are re-used prior to declaring something a “metric”, including points like longevity of data across disciplines and how re-use may or may not reflect the value.

*Premise 5: The WG should continue to exist in maintenance mode or as an IG and work in close coordination with Scholix as usage and citation are closely tied together*

In this WG’s context, dataset usage has been largely focused on views and downloads of research data. Dataset citations could also be understood as a reflection of dataset usage and reusage. Because of this, it’s essential that this WG remain in maintenance mode and work in parallel with the SCHOLIX WG to have a clear message for repositories on how they can uphold best practices for counting, displaying, and aggregating dataset usage and citation. This will be required for there to be a shared understanding going forward of how to meaningfully assess the reach and impact of research data.