IJDC | *Conference Paper*

Metrics to Increase Data Usage Understanding and Transparency

Maria Esteva
University of Texas at Austin

Joshua Freeze
University of Texas at Austin

James Carson
University of Texas at Austin

Craig Jansen
University of Texas at Austin

Abstract

Data metrics are essential to assess the impact of data repositories' holdings and to understand the research practices of the community that they serve. These metrics are useful for reporting to funders, to inform community engagement strategies, and to direct and sustain repository services. In turn, communicating these metrics to the user community conveys transparency and elicits their trust in data sharing. However, because data metrics are time-sensitive and context-dependent, tracking, interpreting, and communicating them is challenging. In this work we introduce data usage analyses including benchmarking and grouping, developed to better assess the impact of the DesignSafe Data Depot, a natural hazards data repository. Make Data Count compliant metrics are analysed in relation to research methods, sub-disciplines, natural hazard types, and time, to learn what data is being used, what influences data usage, and to establish realistic usage expectations. Results are interpreted in relation to the research and publication practices of the community and to natural hazard events. In addition, we introduce strategies to clearly communicate dataset metrics to users.

Submitted 9 February 2024 ~ *Accepted* 20 February 2024

Correspondence should be addressed to Maria Esteva, 3206 South Oak Drive, Austin, TX 78704. Email: maria@tacc.utexas.edu

This paper was presented at the International Digital Curation Conference IDCC24, 19-21 February 2024

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

Data usage and citation metrics are key impact indicators to repositories and users. Repositories use these metrics to understand usage trends, to justify operations, and to improve services. Exposing the metrics publicly conveys transparency and elicits trust in data publication. To data producers, learning that their datasets are used is rewarding. To consumers, it signals the relevance of the datasets (Mayernik et al., 2017).

The Make Data Count (MDC) initiative standardised the tracking and aggregating of data citation and usage, with emphases on the transparency and comparability of metrics across datasets and repositories (Cousijn et al., 2019). Based on best practices included in the COUNTER code of practice for research data, the metrics address the structure and diversity of datasets and how they are used (Fenner et al., 2018) Specifically relevant to this work are metrics on views and downloads, which many repositories expose to the public on the datasets' landing pages.

However, implementing, analysing, and interpreting data usage metrics are not straightforward tasks, as these involve complex conceptual and technical underpinnings. In contrast with citation metrics, the community has not come to a common understanding about their value (van de Sandt et al., 2019). While MDC normalises metrics between large and small datasets, allowing for comparison, contextual elements such as the datasets' domain science, their time of publication, and other events surrounding their use are needed for understanding usage performance. In data catalogues such as DataOne [1] and DataCite Commons [2], as well as in the Zenodo [3] repository, users can see how their metrics evolve over time, but they don't have a way to easily compare their usage to other datasets. In turn, repository administrators and curators need to know how and why data usage behaves over time to direct training and communications, understand gaps in usage, and shape policies. We present work done to address these challenges in the DesignSafe Data Depot Repository (DDR) [4], where MDC metrics were used to learn how usage of different types of data fares in relation to contextual elements present in DDR metadata.

In operation since 2016, DesignSafe is the cyberinfrastructure of the National Science Foundation-funded Natural Hazard Engineering Research Infrastructure. It provides cloud-based tools to manage, analyse, and publish data to understand the impacts of natural hazards. The DDR is the open access repository component, where researchers curate, publish, discover, and access natural hazards research datasets. DDR supports a diverse user community including engineers and social scientists. In engineering, DDR accepts data regarding the impacts of wind, earthquake, storm surge, drought, extreme heat, and wildfires. Social science datasets encompass the study of the human dimensions of natural hazards. As of 2023, DDR is Core Trust Seal Certified [5]. Since 2022, DDR has implemented MDC-compliant metrics, contributing this information to DataCite. In turn, usage metrics such as views and downloads, as well as citations consumed from DataCite, are displayed on the datasets' landing pages.

While the user community has expressed satisfaction about the availability of these metrics, the DDR team sought to make them more useful both to users and to the repository administrators. Data metrics are time-sensitive and context-dependent; when and how much data are published, as well as the data's scientific provenance, all need to be considered when analysing and interpreting them. To provide references for users to understand the degree of usage of their datasets, we devised benchmarks in relation to time elapsed between publication and use, and usage patterns over time were explored through grouping analyses, revealing

[1] DataONE: https://www.dataone.org/
[2] DataCite Commons: https://commons.datacite.org/
[3] Zenodo: https://zenodo.org/
[4] DesignSafe: https://www.designsafe-ci.org/
[5] Core Trust Seal: https://www.coretrustseal.org/

different 'usage bins'. Results were interpreted by a team of curators and domain specialists with knowledge of the research practices and events that trigger usage in the community. Finally, a user interface modal was designed to improve transparency in communicating metrics to users and stakeholders. The different analyses reveal new functions and values of MDC metrics that contribute to understanding data impact and repository transparency.

Related Work

Significant efforts are invested in data repositories, and therefore a system to evaluate their impact is in order (Parr et al., 2019). Data usage, as views and downloads, and data citations, are amongst the metrics used to evaluate impact, and ongoing discussions about their meaning and importance bear on how they are valued by the research community (Pasquetto, Randles, & Borgman, 2017; van de Sandt et al., 2019). While difficult to track and therefore imprecise, data reuse citations are considered very relevant by the scientific community (Kratz & Strasser, 2015), while usage metrics are seen as complementary or of lesser value (Parr et al., 2019). Parallel reuse studies of qualitative data at the UK Data Service and the Finnish Social Science Data Archive include both citations and usage metrics, the latter as number of downloads per dataset (Bishop & Kuula-Luumi, 2017). The study provides comprehensive analyses including reuse by user type, by dataset type, and by reuse purpose based on metadata gathered by the repositories. It also presents various statistical analyses to gauge data reuse over time. In our study we aim to find similar information. Differently, we base our analyses on MDC compliant metrics, and we designed a new set of methods to measure impact.
Christine Borgman argues that data metrics can be used to favour or to play down the resources that they represent and pleads for careful assessments of who benefits and gains with data metrics and how they are obtained (Borgman, 2022). To make them more robust and ensure that they afford fair evaluations metrics must be properly contextualised. To make sense of data metrics, The Meaningful Data Count project proposes a multi-research methods approach including qualitative surveys and bibliometric analyses based on knowledge graphs and DataCite metadata (Ninkov, 2022). The project's goal is to set the path to new data usage and citation analyses, and consequently stimulate scholars to value data publication and reuse. While it introduces different approaches, our work has similar goals. For tighter disciplinary and methodological data usage contextualization, we focus on understanding usage within DDR. Both benchmarking and bin analyses are based on usage data from DDR and complemented with rich metadata that we collect about each published dataset. At this moment, we only analyse data usage metrics, as data citations consumed from DataCite are still sparse. This focused perspective along with the new usage analysis methods that we developed can contribute to existing initiatives to make data metrics more useful and visible.

Data Usage Analyses

We introduce analyses conducted with usage data from January 2022, when MDC metrics were implemented in DDR, through December 2023. The MDC metric used is unique requests (URs), which refers to the number of one-hour web sessions during which a user previewed, downloaded, or copied files associated with a DOI [6]. URs are considered the closest metric for assessing the number of people that have examined the data. Contained within a one-hour

[6] Make Data Count. *Data Usage.* https://makedatacount.org/data-usage/

session, the metric allows fair usage comparison between users completing actions on datasets that have different number of files. URs are the equivalent of downloads for DataCite[7]

URs are analysed in conjunction with metadata gathered in DDR. Throughout the data curation process, authors use controlled terms to label datasets according to scientific domain, research data type, natural hazard type, and other required descriptive elements that we gather. Along with time, these elements are used to enrich the context of usage. Results are interpreted based on lessons learned about research practices, data publication, and data reuse of the community (Borgman, Wallis, & Enyedy, 2006; Esteva & Rathje, 2020). In the following section, the different analyses are presented as a progression, reflecting how we explore usage.

Examining Usage in Context with Research Data Types

To help understand how usage of the different research data types evolves, Figure 1 shows differences in usage behaviours as URs per quarter for different research data types published during the study period. DesignSafe has four categories of research data type that users are required to label their datasets with: Simulations, Experiments, Field Research, and Other. The latter is a broad category that may entail Jupyter notebooks, benchmark data, integrated datasets, and a variety of data reports.

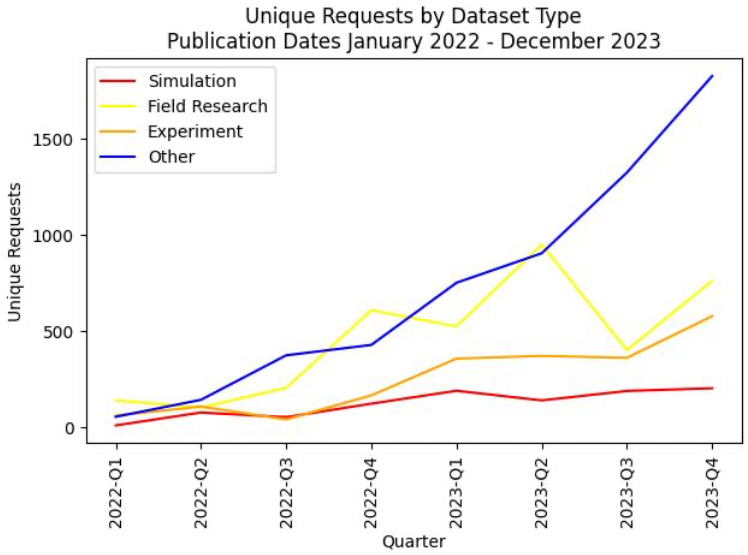

Figure 1. Unique Requests by data type by quarter.

It can be observed that the trend is that usage increases over time, but different research data types show different shapes. Because it aggregates different data resources, "Other" research data types are the most used. Spikes of usage in "Field Research" are motivated by usage of datasets published shortly after earthquakes in Mexico (September 2022) and Turkey (February 2023), suggesting the urgent need for this data after a natural hazard event. In contrast, "Experiment" and "Simulation" datasets are related to unique research themes and generated during lengthy studies. Their usage is less dramatic or predictable, occurring when related research papers are presented at conferences, shared by word of mouth, or found via searches and browsing.

[7] DataCite Support. *Views and Downloads.* https://support.datacite.org/docs/contributing

Benchmarking Usage

The aggregated usage numbers for downloads (DataCite metrics term for URs) presented on the datasets' landing pages do not provide context to gauge usage in comparison to other datasets in the repository. To address these gaps, we explored benchmarking to identify low, medium, and high usage.

For the benchmarks to be realistic and fair, comparisons are made between publicly available datasets within the same domain for which there is usage data for a given period of time. For purposes of developing the methodology, we considered URs for datasets during the 12-month period following their publication. Looking at the distribution of usage data for the period, we observed that it is sparse on the high usage end. This led us to not include datasets with exceptionally high usage of 100 or more URs, which we call outliers. The remaining data is somewhat skewed to the right compared to a normal distribution, showing a slower drop off than the left. However, we treat it as normal in determining benchmark percentages.

From the usage figures, we created a histogram showing the number of datasets for each count of URs (see Figure 2). This resulted in defining medium as usage within one standard deviation of the mean (middle 68.2%). Low and high usage are under and over one standard deviation respectively (bottom and top 15.9%). Results classified low usage for the first 12 months after publication as four or fewer URs, high as 40 and above, and medium as spanning the middle between four and 40. These results are particular to usage in DDR, and other repositories may find different boundaries using this methodology.

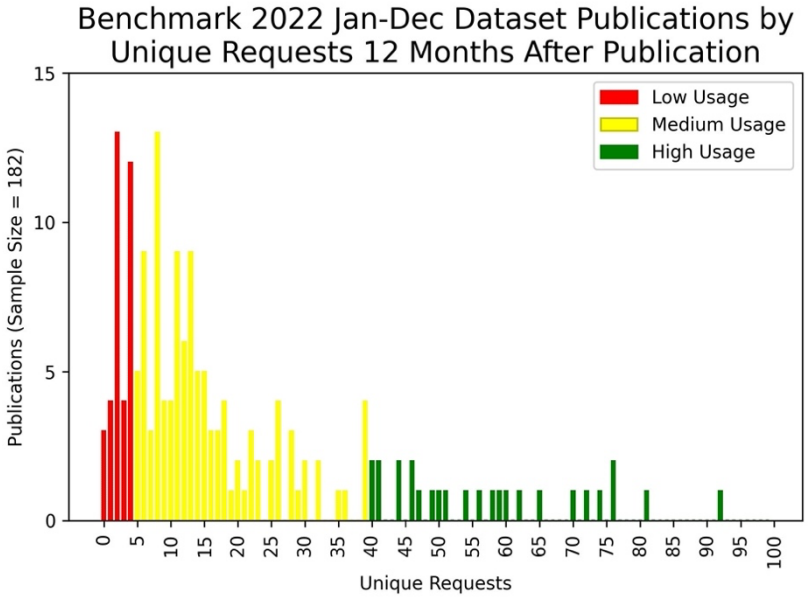

Figure 2. Benchmarking usage: 12 months after publication for 182 datasets.

To provide benchmark information for datasets available for a different time frame, we repeated the calculation for all datasets for which we have 18 months of usage data, those published January–June 2022 (see Figure 3). The upper limit of the low category rose to nine. The boundary between medium and high did not change significantly.

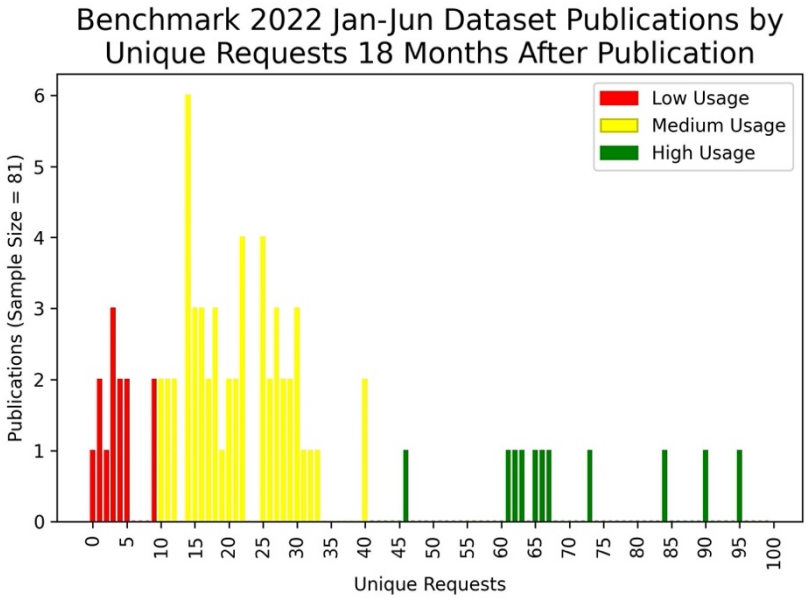

Figure 3. Benchmarking usage: 18 months after publication for 81 datasets.

Because usage changes over time, we will periodically recalculate benchmarks to update boundary values for timeframes of 12, 18, and 24 months. As we gather more usage data for each time frame, we plan to explore benchmarking per research data type.

Outliers: Exceptionally High Usage

Based on the current DDR usage data, we define outliers as those with 100 URs or more, but other repositories may need to define them differently. For 12 and 18 months, data becomes sparse after 92 and 95 URs respectively. We looked for explanations for the outliers. Analysis of the most used dataset (Huang & Swain, 2022) demonstrated that media events surrounding the research featured in the data can substantially increase usage. In Figure 4, peaks coincide with two articles in the New York Times mentioning the research conducted by the authors, the winning of the annual Dataset Award, and publication of a feature about the dataset on DesignSafe's website (August 2022, January 2023, June 2023, respectively). Other instances of exceptionally high usage of datasets are frequently related to the occurrence of a significant natural hazard event (see, for example, in Figure 1 the peaks corresponding to field research).

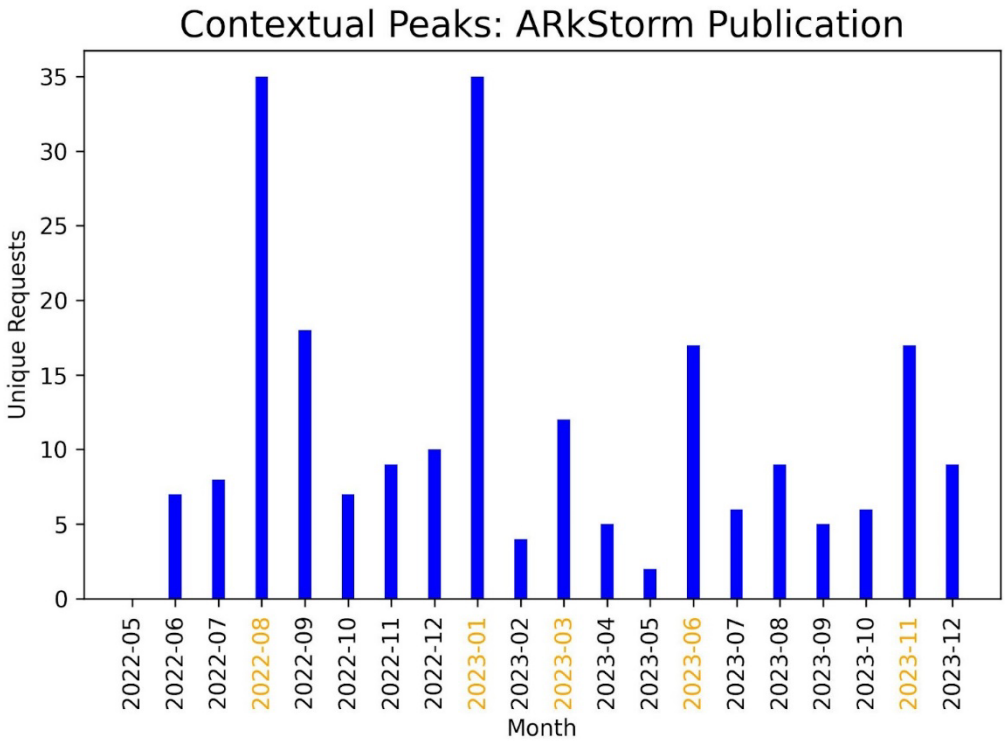

Figure 4. URs per month for Huang & Swain dataset

Grouping Analysis: Usage Bins over Time

Answering questions about how datasets are used over time is key to informing data policies and sustainability strategies. For example, does a dataset's usage sustain over time, or does it diminish following a peak after publication? If so, why and after how long? Are there different trends depending on data types and domains? To answer these questions, we devised criteria to group URs based on common assumptions about data usage that were not yet tested. Examples include that usage peaks soon after publication and quickly fades over time, or that only very few high-profile datasets from seminal research projects or catastrophic natural hazards remain of interest. All these assumptions need to be tested over time.

The methodology that we call "bin analysis" was tested with a cohort of all DDR datasets published between January and June 2022. With a total of 81 dataset publications in the cohort, 36 of them were excluded because they have fewer than 15 total URs during the study period. Such small usage would not lend itself to defining useful patterns or parameters. This left the remaining 45 datasets in the study. For the analysis, usage graphs were plotted for each individual dataset. Because usage is often irregular month-to-month, we chose a period of two months to smooth out this variability. Usage was examined through 15 months after the month of publication, providing eight two-month periods. Usage by period was averaged for all datasets, and datasets were scaled to prevent those with higher usage from dominating the averages. Bins were defined based on two criteria: a) where the first peak is; and b) how long the following tail of significant usage persists. "Early first peak" has a peak in the first two periods. "Long tail" is defined as significant usage occurring in the third or later period following that peak. Significant is defined as 20% of the first peak height but with a minimum of two URs. The criteria generated four bins: early peak, long tail; early peak, short tail; late peak, long tail; and late peak, short tail. A fifth bin type for a very late peak was added to accommodate datasets with their first peak in the final three periods and thus a tail that is not yet established (See Figure 5).

The results show that the 45 datasets fall predominantly into two bins. Nine (20%) have very late peaks and so no categorised tail, but by far the largest group is early peak & long tail, including twenty-seven (60%) datasets. For the remaining three categories, early peak & short tail has four datasets (9%), later peak & long tail has three (7%), and later peak & short tail has two (4%).

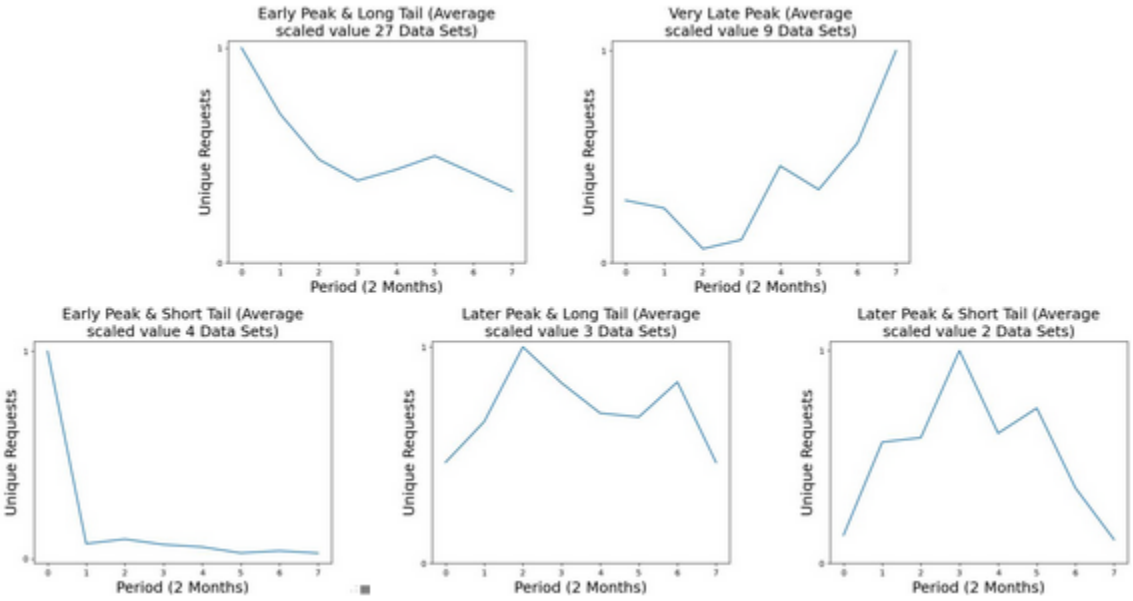

Figure 5. Bin analyses: graphs of average scaled use for each bin. Left to right from top: early peak & long tail; very late peak; early peak & short tail; later peak & long tail; later peak & short tail.

Relations between bin type and research data type, natural hazard type, and domain were also studied with the purpose of better explaining what is being used and how over time. Figure 6 shows that early peak & long tail usage is reasonably balanced between the different research data types. In terms of natural hazard types, we found that although DDR has considerably more earthquake than wind hazard datasets, the bins with long tails are balanced between the two natural hazard types, showing that usage for different hazards is sustained. This result speaks to the interest of the community in the different research methods and natural hazards represented in the collection. We also analysed usage by discipline (Engineering and Social Science) and found that early peak & long tail is notably dominant in the Engineering domain. This may be explained by the fact that natural hazard engineers were, from 2016, the original designated community for DDR, while social scientists joined the community in 2020.

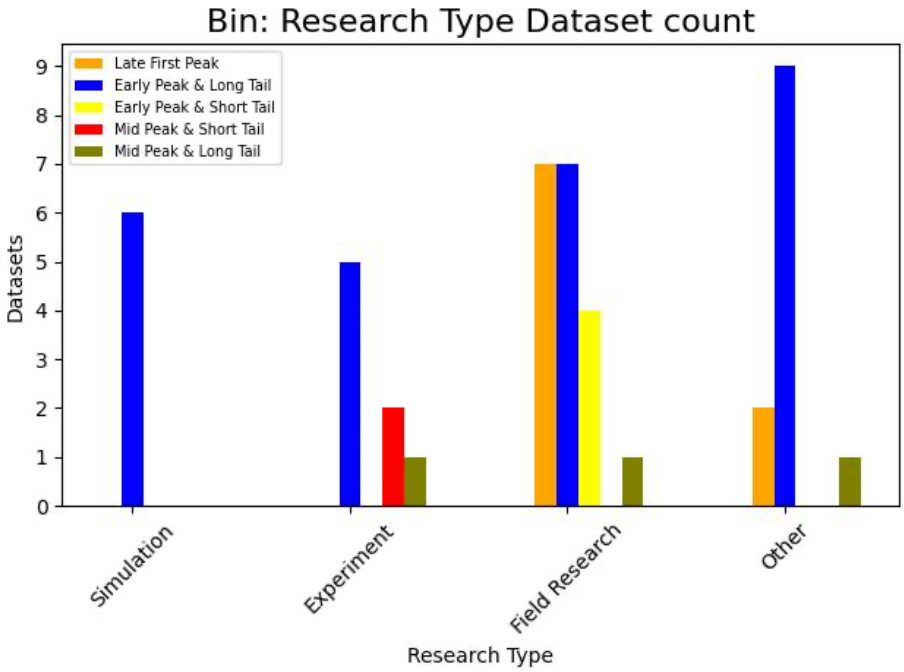

Figure 6. Bin analysis by research data type.

The bin analysis study encompasses a little over a year of usage for six months' worth of dataset publications. While these proportions may evolve, the analyses provide important insights, including that the majority of datasets have sustained usage after a peak that happens upon publication, and that balanced attention is received across the different natural hazard types.

Communicating Metrics

Communicating metrics with clarity is important to achieve users' trust. To complement the usage figures for views and downloads provided for each dataset publication, we designed a modal window that explains the meaning of the different MDC metrics, including their relation to views and downloads. Metrics are then presented by year and quarter to allow users to track usage across time (see Figure 7). Also in the modal window is a link to the Data Metrics section in our User Guide for more detailed information about the MDC metrics and how those are recorded in the repository.

Along with documentation explaining how they are calculated, we plan to make benchmarks available in the Data Metrics section so users can verify how a given dataset performs in relation to others in DDR within the same timeframe after publication. Moving forward, as we collect and analyse more usage data over time, bin analyses may be relevant for users' consumption.

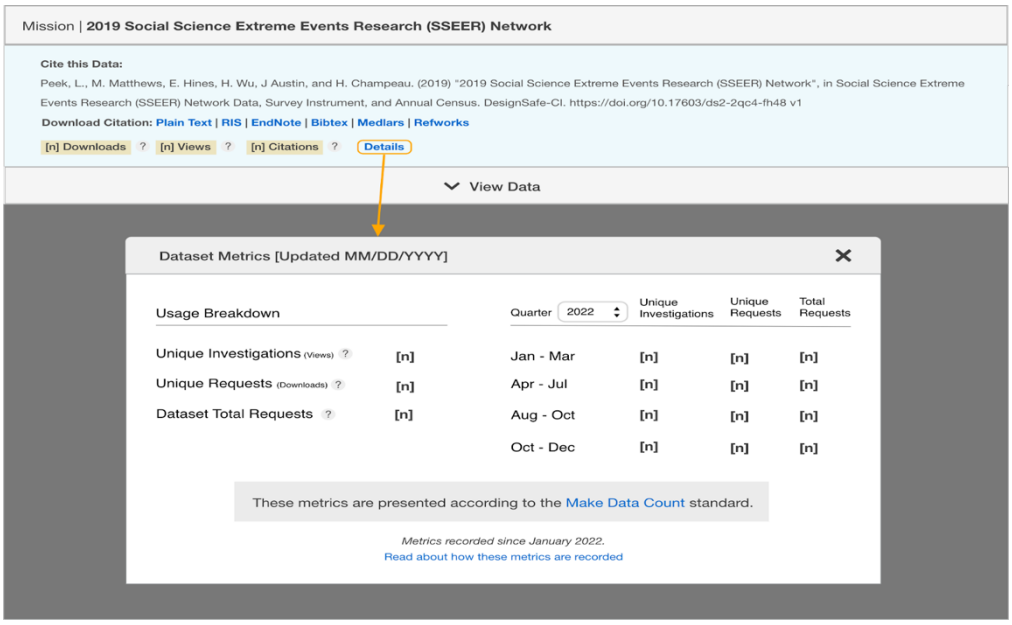

Figure 7. Modal implemented to explain usage metrics.

Conclusions

We conducted MDC metrics analyses to explore patterns of usage over time for the different communities served by DDR and the hazards that they study. The work suggests that data usage is a complex phenomenon that needs to be continuously studied from different perspectives.

The research and publication practices of the different communities have a bearing on what types of datasets are more used, and the occurrence of natural hazards and research dissemination are major usage boosters. Benchmarking conducted within DDR can help researchers gauge usage expectations realistically. And yet, low or medium benchmarking results should not be discouraging. Instead, they should stimulate data publishers and data curators to further investigate whether and how datasets can be improved or promoted for further usage. Bin analysis criteria was designed based on unconfirmed assumptions about data usage over time. We saw that for the data available, the majority of datasets had long tails. Over time and with more testing, we may see that the assumption that most usage vanishes with time was correct, but that our estimate of when that drop-off occurs was too soon. As more datasets have sufficient time elapsed after publication, we may add more bin criteria, such as high total usage during the period analysed.

This first year-and-a-half of metrics constitute a baseline that we will continue evaluating as new users, publications, data types, and events come on board, and more time elapses since we first measure URs. In addition, we will seek feedback on our methods of communicating usage information to learn how users understand, follow, and use the different metrics.

Acknowledgements

This research was supported by the US National Science Foundation DesignSafe Cyberinfrastructure Award #2022469.

References

Bishop, L. & Kuula-Luumi, A. (2017). Revisiting Qualitative Data Reuse: A Decade On. *SAGE Open,* 19,1, https://doi.org/10.1177/2158244016685136

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2006). Building digital libraries for scientific data: An exploratory study of data practices in habitat ecology. In J. Gonzalo, C. Thanos, M. F. Verdejo, & R. C. Carrasco (Eds.), *Lecture Notes in Computer Science: Vol. 4172. Research and Advanced Technology for Digital Libraries* (pp. 170–183). https://doi.org/10.1007/11863878_15

Borgman, C. L. (2022). *Meaningful Data Metrics for Whom?* [presentation slides]. eScholarship. https://escholarship.org/uc/item/593121bt

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal 18*, Article 9. https://doi.org/10.5334/dsj-2019-009

Esteva M., Rathje E. (2020). Perspectives from Data Reuse in the Field of Natural Hazards Engineering. 12th Qualitative and Quantitative Methods in Libraries International Conference (QQML 2020). http://qqml.org/wp-content/uploads/2017/09/Book-of-Abstracts-26-5-2020-.pdf

Fenner, M., Lowenberg D., Jones, M., Needham P., Vieglais D., Abrams S., Cruse P., Chodacki J., (2018). Code of Practice for Research Data Usage Metrics Release 1. PeerJPreprints 6:e26505v1 https://doi.org/10.7287/peerj.preprints.26505v1

Huang, X., & Swain, D. (2022). ARkStorm 2.0: Atmospheric Simulations Depicting Extreme Storm Scenarios Capable of Producing a California Megaflood. *DesignSafe-CI.* https://doi.org/10.17603/ds2-mzgn-cy51

Kratz, J. E., & Strasser, C. (2015). Researcher Perspectives on Publication and Peer Review of Data. *PLOS ONE 10*(2), Article e0117619. https://doi.org/10.1371/journal.pone.0117619

Mayernik, M. S., Hart, D. L., Maull, K. E., & Weber, N. M. (2017). Assessing and Tracing the Outcomes and Impact of Research Infrastructures. *Journal of the Association for Information Science and Technology 68*, 1341–1359. https://doi.org/10.1002/asi.23721

Ninkov, A. (2022). *Accessing, Analyzing and Visualizing Research Data Metadata Using DataCite and Jupyter Notebooks* [presentation slides]. Zenodo. https://doi.org/10.5281/zenodo.6564424

Parr, C., Gries, C., O'Brien, M., Downs, R. R., Duerr, R., Koskela, R., Tarrant, P., Mauli, K. E., Hoebelheinrich, N., & Stall, S. (2019). A Discussion of Value Metrics for Data Repositories in Earth and Environmental Sciences. *Data Science Journal 18*, Article 58. https://doi.org/10.5334/dsj-2019-058

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017) On the Reuse of Scientific Data. *Data Science Journal 16*, Article 8. https://doi.org/10.5334/dsj-2017-008

van de Sandt, S., Dallmeier–Tiessen, S., Petras, V., & Lavasa, A. (2019). The Definition of Reuse. *Data Science Journal 18,* Article 22. https://doi.org/10.5334/dsj-2019-022