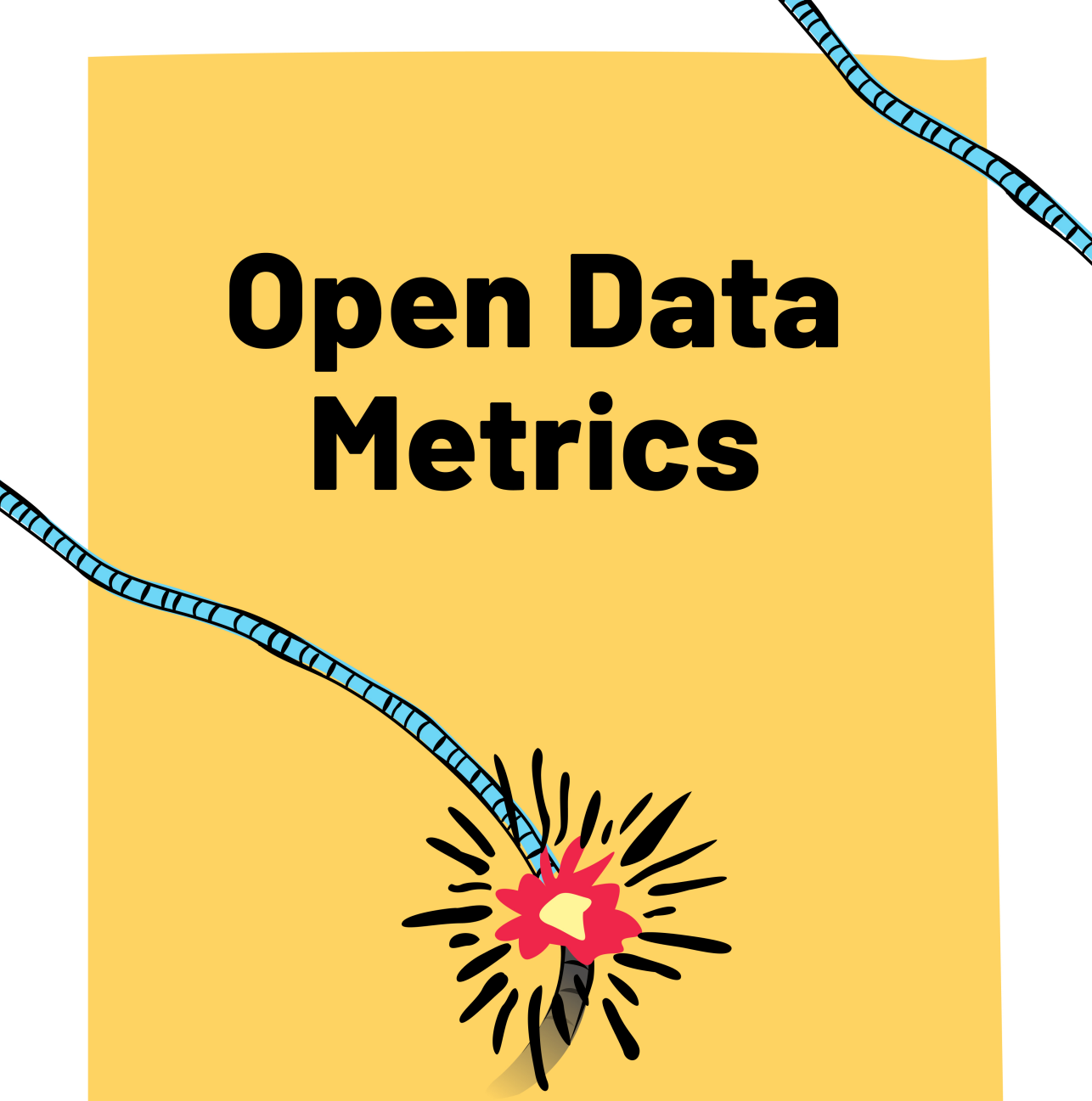
Open Data Metrics



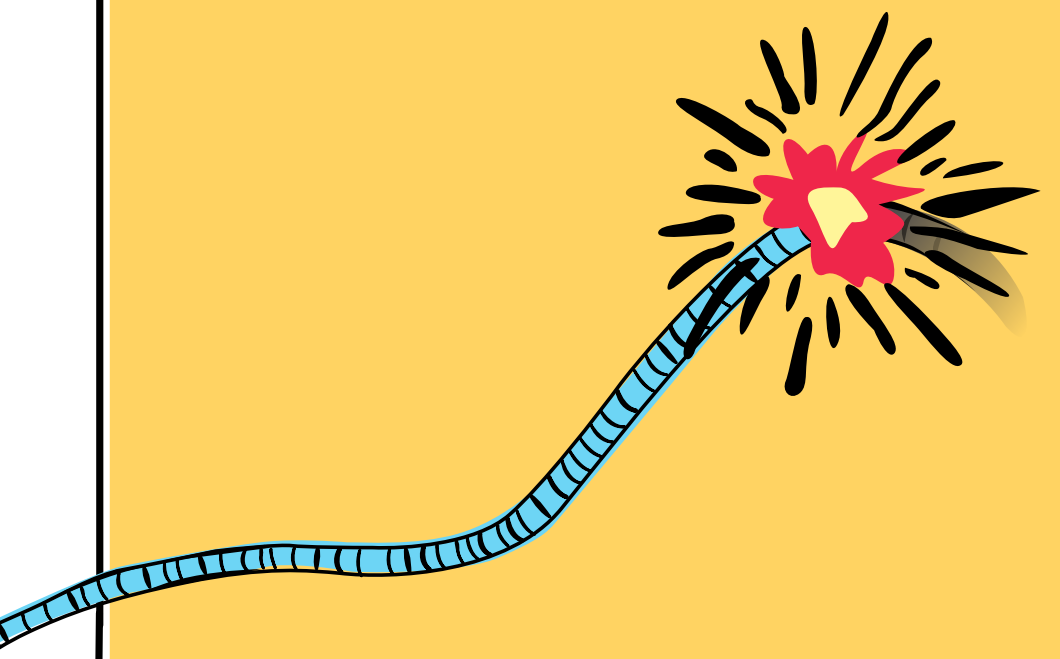
Lighting the Fire

Daniella Lowenberg • John Chodacki • Martin Fenner
Jennifer Kemp • Matthew Jones

Table of Contents

"*A book, too, can be a star, a living fire to lighten the darkness, leading out into the expanding universe."* Madeleine L'Engle

1 Building a case for open data metrics



Ask research communities to ponder whether, given a choice, in a hundred years they would rather have access to all the published papers from a given year, or all the data collected in that year, and, invariably, they would have to think that through. Eventually they will concede that, except for a possible handful of notable papers in their field, research data are the more valuable of the two. This is a remarkable acknowledgement of the power of data and a conspicuous indictment of many of the research supporting communities that have not yet invested in the necessary systems to support research data sharing.

The reality that research data are so foundational can be seen in the tradition of researchers citing data as sources. In one recent analysis in oceanography, Belter (2014) explored three key datasets that have been cited in thousands of scientific papers. If they were viewed as papers, "each of the three datasets would be ranked in the top 1% for citation counts of all articles published in Oceanography during the same year, while [two] would be ranked in the top 0.1%." [1] While not all research datasets are this large, complex, or powerful, this example directly evidences the need to amalgamate similar data into something more valuable than a paper for the scientific community.

Thousands of papers from national synthesis centers [2] that reuse data for cross-scale comparisons [3] are evidence of widespread data reuse and the benefits of sharing open data. For example, with over 21,000 citations, Constanza et al. (1997) is one of the top cited papers of all time and was completely based on the reuse of existing data about the economic value of natural ecosystems [4] — and it helped catalyze the formation of the new subdiscipline of ecological economics.

Open data sharing and the publishing of data have gained widespread acceptance [5] (even with warnings that data publication may not be the

right concept for making research data open[6]). Researchers cite, share, and reuse data, yet little information on this data ecosystem is collected, let alone in a structured, standardized, or accessible way. This points to a gaping hole in the understanding of the impact of data on research, policy, management, and society.

Notes

1. Belter, C. W. (2014). Measuring the Value of Research Data: A Citation Analysis of Oceanographic Data Sets. *PLOS ONE*, *9*(3), e92590. https://doi.org/10.1371/journal.pone.0092590

2. Baron, J. S., Specht, A., Garnier, E., Bishop, P., Campbell, C. A., Davis, F. W., … Winter, M. (2017). Synthesis Centers as Critical Research Infrastructure. *BioScience*, *67*(8), 750–759. https://doi.org/10.1093/biosci/bix053

3. Specht, A., Gordon, I. J., Lambers, H., Phinn, S. R., & Phinn, S. R. (2015). Catalysing transdisciplinary synthesis in ecosystem science and management. *Science of The Total Environment*, *534*, 1–3. https://doi.org/10.1016/j.scitotenv.2015.06.044

4. Costanza, R., d'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., … van den Belt, M. (1997). The value of the world's ecosystem services and natural capital. *Nature*, *387*(6630), 253–260. https://doi.org/10.1038/387253a0

5. Kratz, J., & Strasser, C. (2014). Data publication consensus and controversies. *F1000Research*, *3*(94), 1–18. https://doi.org/10.12688/f1000research.4518

6. Parsons, M. A., & Fox, P. A. (2013). Is Data Publication the Right Metaphor? *Data Science Journal*, *12*, WDS32–WDS46. https://doi.org/10.2481/dsj.WDS-042

Why open?

Historically, many fields of research have been conducted utilizing closed platforms and technologies. While there are many reasons for this phenomenon, it has roots in both the technical (lack of capabilities or technology) and the cultural (pressures and incentives to be the first to publish). Furthermore, there has also been a perception in some fields that collaborative work would be less productive than work conducted independently. To further this trend, most scientific outputs were also published in closed journal platforms which had researcher findings trapped within closed platforms or behind paywalls.

Still, views in research are changing, and the default in many fields is beginning to shift to be far more open. The *open source* movement paved the way by normalizing the idea that free and open software had value. The idea of open source content originated in the hacker culture of the 70s and 80s and was truly launched by the formation of the GNU Project in 1983,[1] leading to the Free Software Foundation shortly thereafter. At the time, most software was proprietary and restricted by commercial use licenses. The radical idea that people might simply give away such a valuable commodity met resistance throughout the software world. Nevertheless, the power of the open approach was reiterated as the World Wide Web took shape in the mid-90s, resulting in the widespread acceptance of the value of being open in a traditionally closed world.

Open access, a movement that gained traction in the 90s, was built on the momentum from open source and proposed that unrestricted

access to research articles would speed up scientific progress. Now, a few decades on, open access demands, at a minimum, universal access to scholarly content without a fee charged to readers, and with a license that allows reuse and redistribution. As with most movements, there is some disagreement about terminology, approaches, and 'how far' to take the basic idea. Still, despite complicated disagreements, and the persistence of traditional, subscription-based publishing, the open access movement has been a landmark in changing the culture of science.

Following on these successes, and acknowledging the inherent value that research data have, as a foundation of science, the *open data* movement has the goal to make research data accessible, reusable, and distributable, and has the potential to revolutionize scientific discovery.

Even though there is no 'data or perish' equivalent to 'publish or perish', [2] the open data movement has struggled with participation, for reasons both technical and cultural. While many initiatives, stakeholders, and services have engaged and built support for the open data movement, there is an explicit need for open metrics to evaluate and demonstrate research data value.

The goal of open science practices is to make research more transparent and accountable through retooling for the research process. Since the evaluation and analysis *of* research are both inherent *to* and inherent *in* research, it only makes sense that these would also be part and parcel of open science. Exposing these analyses both establishes the practice as a norm and functions to share findings with the research community. This extension of the established scientific method is a good example of the kind of work that both reflects and supports the goals of open science.

Notes

1. GNU — Initial Announcement. (1983, September 27). Retrieved November 1, 2019, from https://www.gnu.org/gnu/initial-announcement.html
2. Colloquial term referring to the pressure for academic scientists to publish papers for the advancements of their careers

Igniting community change

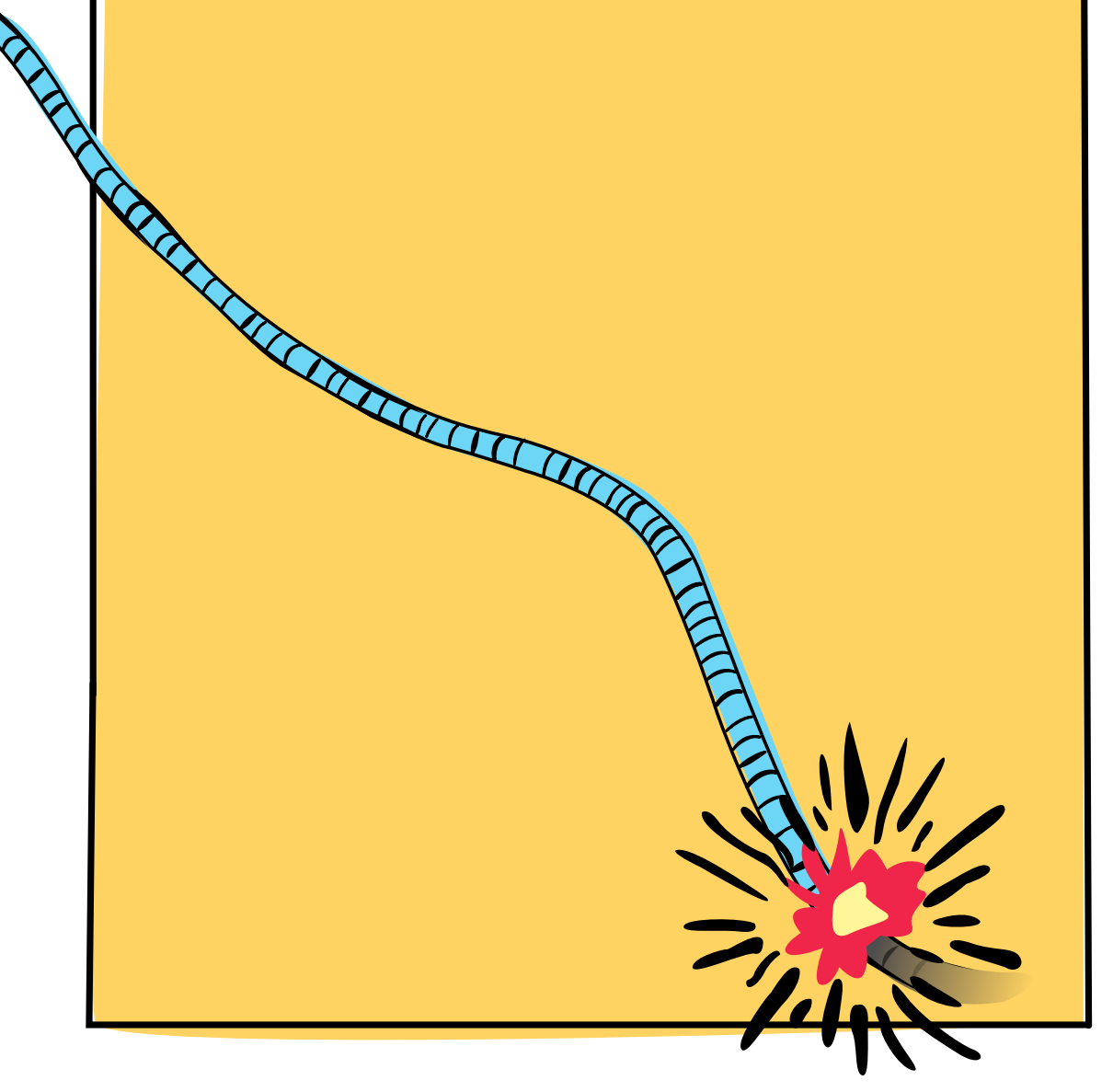
When faced with a systematic change as large and complex as open data metrics, it can be hard to see how any individual, or indeed even large organizations, can begin a transformation. The sheer scale of it all can be paralyzing. Nevertheless, besides the value in being able to assess the impact of research data, there is a strong differentiator in favor of open data metrics: there is no existing, monolithic system to overhaul. These metrics are being implemented from scratch, which is why community discussion and participation are so necessary. In light of this, every interested player is given the opportunity to effect change. Given the existing awareness in the community of the value of data sharing, stakeholder contributions, incrementally or larger-scale, can shape and drive the future of data metrics.

Our intention is to light a fire in the community so discussion and data metrics development can progress. This book describes current progress in standards and infrastructures and highlights how a bright future of open data metrics depends on community participation from researchers and *research supporters:*[1] repositories, publishers, funders, and institutions.

Notes

1. Chodacki, J., Cruse, P., Lin, J., Neylon, C., Pattinson, D., & Strasser, C. (2018). *Supporting Research Communications: A guide*. https://doi.org/10.5281/ZENODO.3524663

2 Understanding data and data metrics



Data has always been the basis of research, and researchers have long treated it as such. In recent decades, the growth of digital communications has facilitated the discoverability and reuse of data. Research supporting communities are beginning to understand that publishing, citing, reuse and tracking of research data are key to supporting open science and a fuller, persistent, scholarly record. Data are now commonly deposited in repositories of various kinds for reuse and preservation, assigned identifiers and described in distributed metadata for integration into systems and services throughout scholarly communications. In short, research data are now increasingly recognized, on their own and along with the related publications, as having their own inherent value and thus being in need of their own standardized treatments of citations and associated metrics.

Our starting point

Research supporters are interested in understanding the impact and reach of shared data. The natural inclination is to evaluate this impact with metrics such as views, downloads, and citations in exactly the ways we have done with articles. Taking this approach, however, conflates many issues and perpetuates preconceived notions that articles and data should be treated as if they are the same. Of course, multiple outputs may be from the same research project or tightly correlated, and we may need to rely on similar approaches to counting usage and citation, but we cannot immediately assign the same frameworks to data that we have accepted for articles. [1]

As in any field or discipline, similar words or phrases can be assumed to have specific meaning but may be interpreted differently by our peers. Within communications communities, *metrics* is an especially

loaded term. Therefore, it's worth describing what is meant in this book when we use the term.

Metrics inherently impart some level of assessment, and, by design, they are intended to communicate value. Metrics are contextual in that they must be interpreted within their context. Though they are often thought of as a number, the combination and degree of selection and processing that goes into metrics sets them apart from raw counts.

As much as adopted metrics may be (or appear to be) thoughtful and considered, vetted by experts, transparent, and intended to enlighten, there is no number, grade, or report that can be taken entirely at face value. What appears to be a time saver is really a cleverly disguised invitation to a behind-the-scenes trove of details that can be rewarding or at least illuminating. Responsible metrics providers understand how a particular metric was calculated and with what data.

Therefore, before getting to a point of proper data metrics, it's necessary to outline basic terms and states of practice. What follows are the definitions, proposed framework, and necessary steps, for the community to consider in order to achieve truly open and reliable data metrics.

Notes

1. It is well understood that *metrics* include more than citations and usage. Additional attention is needed regarding altmetrics for research data, for example, mentions on Twitter and Wikipedia. However, as Kratz, J., & Strasser, C. (2015) concluded, our first step towards data metrics should focus on data citation and data usage. Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific Data*, *2*, 150039. https://doi.org/10.1038/sdata.2015.39

Many ways to define a dataset

There are different ways of defining what constitutes a *dataset*, and both the researchers and those that support the research build their definitions on several factors. These factors include discipline-specific characteristics, the structure of the data, or various logical approaches. These factors influence how data is modeled, stored, and accessed which in turn affects how it is counted, reported, and assessed. While there are significant variations in approaches, we can still converge into actionable frameworks for capturing reliable data metrics.

Disciplines as diverse as human cultural ethnography, atmospheric chemistry, terrestrial ecology, and environmental microbiology produce data that are stored and preserved in regional repositories such as the NSF Arctic Data Center [1] or in domain-agnostic repositories such as Dryad. [2] Each subdiscipline collects different types of data, ranging from audio interviews with human subjects, to sensor data streams measuring atmospheric gas concentrations, to gene sequence data from water samples. Moreover, each conceptualizes their datasets differently. For the ethnographer, the logical unit of data might be a set of interview responses to a single survey instrument while the chemist might generate gigabytes of data daily and would choose to publish their data by month and region. The ethnographer would likely assign a single citable identifier such as a DOI (Digital Object Identifier) to the dataset, while a life science researcher such as a geneticist would assign an accession number to each genetic sequence. Thus, the size and granularity of each dataset and the segmentation of the data into

identifiable dataset units will vary across these disciplines, which in turn will affect how usage and citation counts are aggregated.

Researchers also naturally choose different organizational structures for their data. Data generated from experiments often have a logical sampling unit and set of treatments that lead to clear boundaries for publishing the data as a dataset. In contrast, observational data might consist of a high frequency of data spanning multiyear time periods and large spatial regions. These are often segmented into multiple datasets along such parameters for ease of management. These choices can determine both the types of identifiers assigned and the granularity of reporting for citation and usage counts.

Even when the overall extent of a dataset is clear, there are many ways to organize and represent the components of the dataset into a coherent set of files containing tables, text, images, and other media formats. For example, a single geospatial dataset with weather measurements at different locations and with different timestamps could be organized in different ways:

- As a table of time series measurements, with one row for each sampled location and time, placing each measured parameter such as wind speed or temperature in a separate file, or as multiple columns in the same file.

- As the entire time series, stored in a single table. However, if it is large, many researchers would break the table up into multiple identically structured tables for each month, year, or spatial region.

- As a spatial vector image with point features for the spatial locations, and attributes of those points containing the time series.

— Or, as a grid where some researchers might choose to organize the data as a raster image in which the value of each parameter is the cell value in a three-dimensional matrix of x, y, and time, and with one matrix for each parameter in the dataset.

Each of these organizational approaches means that the same data could be represented in differing numbers of files, often each with their own identifier, and that these files could be aggregated into different numbers of datasets for easier management. This in turn could affect how usage and citation counts are interpreted.

A conceptual model for normalizing approaches

The many ways to organize data lead to vast differences in interpretation of usage and citation statistics, defined in the next section. As a community, we need working definitions of the standard concepts regarding the aggregation of data into composite datasets. Such definitions can provide the terminology needed for repositories to consistently organize their data holdings, as well as a basis for consistent reporting for usage and citation counts that are described in the "Coordinating emerging tools and standards" chapter.

For decades, the COUNTER Code of Practice [3] has defined standards for reporting statistics for scholarly literature, as well as tools for consistent and comparable reporting. Recognizing the need for a similar level of community agreement on reporting standards for data, the Make Data Count [4] initiative worked with a community of repository stakeholders to define the COUNTER Code of Practice for

Research Data.[5] By defining key terms and methods of consistently applying them to data, this approach allows for addressing the differences in how communities model data and thereby report on usage statistics.

The core concepts in the COUNTER Code of Practice for Research Data standard centers around defining:

- *datasets,* defined as "an aggregation of data, published or curated by a single agent, and available for access or download in one or more formats, with accompanying metadata."

- *components* each of which is "part of the data available for a dataset that can be accessed or downloaded individually."

- *versions* which represent "significant changes to the content and/or metadata, associated with changes in one or more components, and that would result in changes to fixity attributes of the components."

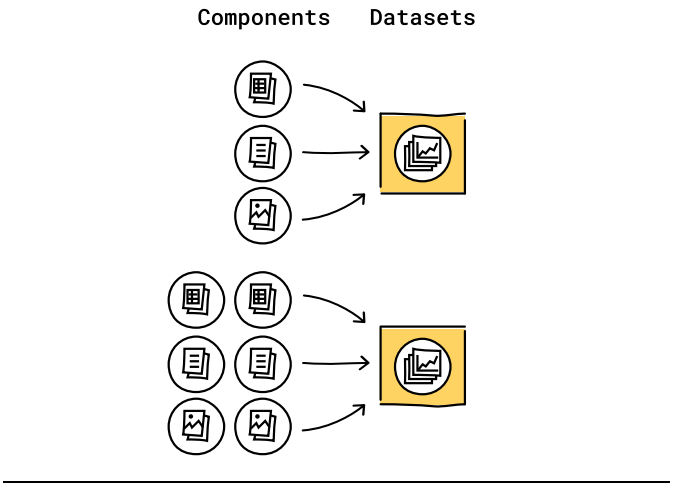
Each of these concepts is tied to related concepts in COUNTER version 5 (2019), clearly establishing the ways in which reporting units for data are similar to and different from the reporting units for articles. Though these terms are jargon heavy, being explicit about the terminology allows for consistency of reporting. Repositories that might normally label their data as a database or a data table can map their approach to the standard dataset and component terminology provided by COUNTER, providing a degree of consistency that is critical for reporting.

Evolving standards to capture dataset complexities

While the current version of the COUNTER Code of Practice for Research Data is a good starting point there is additional work to be done to address the complexities of research data. There are subtleties, inherent in the structure of a dataset, that need to be acknowledged. While the current version offers a structure to dataset usage for the community to get started with, additional work will be required in future versions of the standard to support open data metrics.

Granularity

Many research communities have normalized the practices regarding the level of *granularity* they use to package datasets. However, there can be variation in practice and limitations to these approaches. When community norms do exist, they can be based on tradition as well as practical considerations such as file sizes and file types that are common to that community. As discussed earlier, a single research group or even a single researcher can, nevertheless, package their data in many ways (as tables, raster images, or vector graphics, etc.).

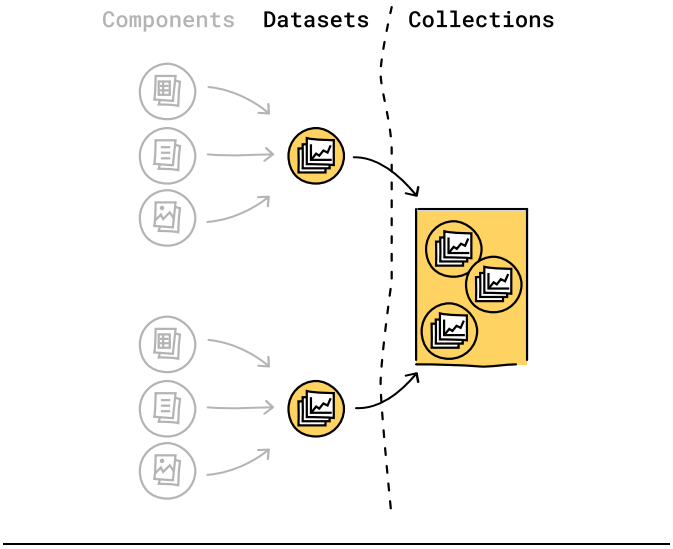


Dataset granularity

Collections

Datasets are not only made available at different levels of granularity but can also be grouped together in more than one way, via *collections*. This grouping is frequently used for long-running projects in which, for example, an annual dataset is collected and then each year's data are closely related and contained within an umbrella project with consistent methods. This becomes complicated when the same datasets are part of multiple collections.

Datasets can then be viewed, downloaded, or cited in multiple ways, making tracking of data metrics across collections complex. For example, a researcher might view datasets individually, or as part of a collection, and then choose to download individual components from the collection view. In many ways, this situation is analogous to the relationship between data components and datasets, but at a higher level in the organizational hierarchy. As with components, community size, and data sharing practices, other factors also play an important role.



Versions

Practices for storage and sharing of *versions* of datasets vary widely. Ultimately, it is the decision of the data repository on how versioning is handled. Versioning of data is important for specificity and verifiability but can be a challenge for data metrics. The challenge for versioned data is best addressed by using identifiers for each version and linking these versions via metadata. Aggregation using identifiers allows for summary data metrics for all versions. The THOR project [6] has articulated best practices for data versioning (see box below).

Practices for Data Versioning

1. Major version changes require a new persistent identifier and new set of metadata, whereas for minor version changes only the data and/or metadata are updated; the persistent identifier does not change

2. A naming convention for the persistent identifier should not be the only place where version information is encoded

3. Both the version number and related identifiers of other versions can be described in the metadata

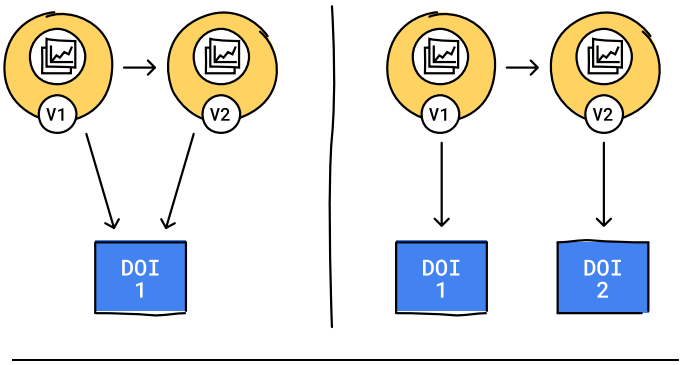
4. Both the version number and related identifiers of other versions can be included in the landing page

5. Humans and machines should be able to easily see multiple versions if they exist, and be able to tell whether they are looking at the newest version of a dataset

6. Data and metadata of older versions should be kept available if possible, using a tombstone page if the data are no longer available

7. Information about what changed in comparison to the previous version is desirable.

8. A collection that includes all versions of a dataset can be assigned a persistent identifier and aggregate their version information



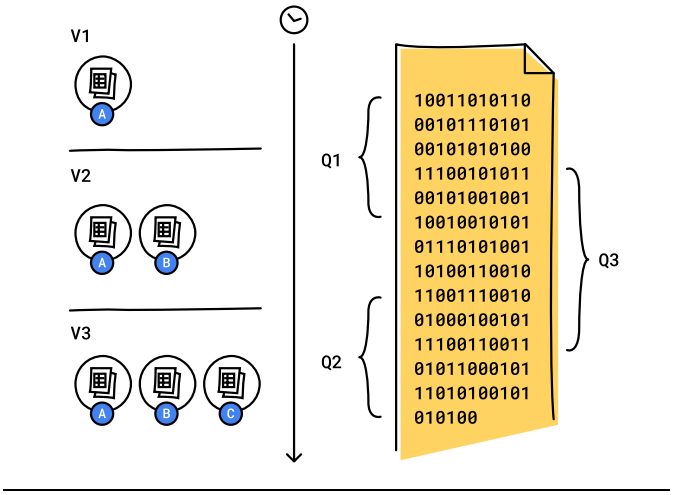
Versioned data

Dynamic data

Dynamic or *evolving data* is an extreme case of versioned data where datasets are not collections of discrete data components, but instead act more like databases that are constantly being fed new information. Much of the archival community focuses on discrete data snapshots that can be assigned an identifier and archived as a unit.[7] In the dynamic data model, data are incrementally appended onto the end of a database, often using time stamps as part of the primary key for data access. In some observational use cases, sensors can generate hundreds or thousands of new data records every second (e.g., high frequency radar data), while also being operated over years or decades.

As shown in the dynamic data figure below, the evolution of a dynamic dataset evolves over time. Discretely identified snapshots of data can be added in each version without affecting what was previously added. People can access each of the discrete components, and users who access each component will always get the same data content. However, in the dynamic data case, the data are also appended to a continuously growing database. New records arrive so rapidly that users who access the data do so with custom queries, each of which involves downloading subsets of the data (which can be overlapping sets of records), timestamping the query, and assigning a persistent identifier to it.

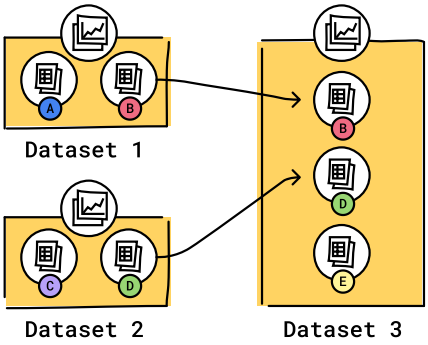
Many complications come into play when we try to impose the versioning concept onto a dynamic dataset. While we can aggregate metrics using a collection identifier to access summary data metrics for all components, the dynamic datasets are constantly evolving. So, this approach is fundamentally different from the collections and versioning examples described previously.



Dynamic data

Derived data

Researchers frequently combine, split, and transform a set of source datasets; we can refer to this new dataset as a *derived dataset*. A derived dataset is like a new version of a dataset, with the critical but subtle distinction that the derived dataset is both fundamentally transformed from the original, thereby making it new. The derived dataset also incorporates data from the original source datasets, so the dataset is not independent of the original, and contains new information, so it is not just a new version.



Derived data

In the example shown in the figure above, the derived dataset has been created by integrating some data from components from two different source datasets, with some new data in the component. However, there are other components that were not part of the source datasets.

To accurately reflect these relationships, we will need a community-sanctioned mechanism to weight the relationship between derived data and source data. One implementation that is available today is connecting data usage and data citation for derived datasets back to the source datasets via their connection in metadata and using the DataCite PID Graph service [8] for this. In the longer run, the community as a whole can continue to discuss the issue of transitive credit for creators of source datasets when derived datasets are downloaded and cited.

Data citation and *data usage* information obtained may not be easily comparable across communities because of these differences in practices around data granularity. Community size, data sharing practices, and other factors play an important role. While these topic areas need some level of community input and development, they do not hinder work that can be done on standardizing approaches to counting and evaluating research data.

Notes

1. NSF Arctic Data Center — The primary data and software repository for NSF Arctic research. (2019). Retrieved November 1, 2019, from https://arcticdata.io/

2. Dryad — Publish and Preserve your Data. (2019). Retrieved November 1, 2019, from https://datadryad.org/stash

3. COUNTER Code of Practice Release 5. (2017, July 1). Retrieved October 29, 2019, from https://www.projectcounter.org/code-of-practice-five-sections/abstract/

4. Make Data Count. (2017). Retrieved November 1, 2019, from https://makedatacount.org/

5. Fenner, M., Lowenberg, D., Jones, M. B., Needham, P., Vieglais, D., Abrams, S., … Chodacki, J. (2018). The COUNTER Code of Practice for Research Data. *Project Counter*. Retrieved from https://www.projectcounter.org/code-of-practice-rd-sections/foreword/

6. Project THOR — Technical and Human infrastructure for Open Research. (2015). Retrieved November 1, 2019, from Project THOR website: https://project-thor.eu/

7. Open Archives Initiative. (2019). Retrieved November 1, 2019, from https://www.openarchives.org/

8. Fenner, M. (2019). The DataCite GraphQL API is now open for (pre-release) business. *DataCite Blog*. https://doi.org/10.5438/QAB1-N315

What we mean by data usage and data citation

Two popular evaluations of use and reuse of data are *usage* and *citation.* Data citation may be interpreted as a count of usage, but the two should be defined separately.

Data usage

Data usage is counted as the accesses of a dataset or its associated metadata. This can be defined as *views* (for example, metadata, 3D models, images displayed on the landing page) and *downloads* (file level or dataset level). However, without a standard for these counts, the definitions of views and downloads used by various stakeholders have been arbitrary. As a result, we see significant variety in how repositories both count and display views and downloads. Currently, to compare the downloads across datasets within a repository, or across repositories, would be comparing apples to oranges, as we do not know where these numbers are derived from, nor exactly what they apply to.

Counts are not themselves metrics, but to get to the point where data metrics can be meaningfully derived, it is essential that counting procedures at repositories are standardized. The COUNTER Code of Practice for Research Data, introduced in the last section, is a standard

for counting data usage, but there is not yet a standard around whether to display counts. Repositories may choose whether to expose these numbers for a variety of reasons. For example, counts may carry weight about the perceived importance or impact of the repository and the datasets that it hosts.

An example of usage and the effort for repositories to support it may be useful. Usage tracking at the National Science Foundation (NSF) Arctic Data Center showed a massive spike in views and downloads for an Arctic sea ice dataset in 2015 (see Example DataONE display figure in the "Coordinating emerging tools and standards" chapter). Upon investigation, the repository was able to determine that the usage spike originated from a large class on machine learning in computer science where many students downloaded the same large dataset for a class exercise. While the spike appeared anomalous, it in fact reflected an important reuse of the data for education. Data uses for teaching, for policy applications, for management, and for outreach and engagement may rarely result in a citation but are strong signals in data usage statistics.

Data usage provide statistics about interests in datasets and may also reflect use and access to data by groups that don't usually publish in academic publications. These statistics have value to the broader community in understanding reuse and linking it to trends in discovery, access, and citation.

Data citation

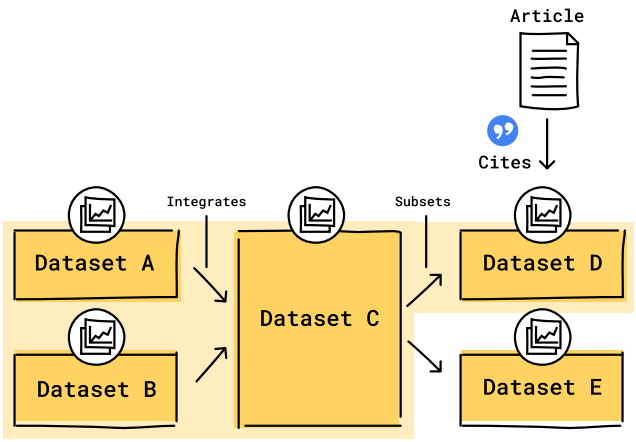
Data citation is commonly thought of as the traditional practice of one article citing another. However, datasets differ from articles in their structure, content, and use, and we need to think about data citations more broadly as recognition linkages between data and other outputs.

One way to consider data citations is articles citing data, where authors cite a dataset alongside the other articles that they used to inform their research. Some members of the community, including groups such as Scholix, have been campaigning for increased adoption of these citations, suggesting they are analogous to article citations. However, the research community has not yet commonly adopted available data citation best practices. Citing data alone isn't enough though, since citations and their associated metadata must propagate through various workflows and systems such as submission platforms and metadata vendors, and these citations are often lost along the way to Crossref and other indexing services. For example, data citations are often removed from article metadata before it is delivered to Crossref. Until this paradigm changes, the ability to cite data in many journal articles is lost. This in turn inhibits both discoverability and counting of such citations. The practice of citing data in articles and indexing these citations within open frameworks requires further adoption and understanding of workflow and other issues to make the process more efficient and effective.

The other way to consider data citations is data citing data. This needs to reflect the complexity of datasets that may include collections,

versions and/or derived data as discussed in the previous chapter. In this context, there may be no associated article published.

We can define this as data citation in the provenance sense. Whenever a researcher works with and uses data, that data is included in their computational workflows, usually multi-step data processing. The flow of data from one step to another, and out of one software package and into another, can produce many intermediate datasets that are archived and identified independently before the final dataset. The references between the steps are themselves a form of data citation, and explicitly represent the information needed to understand the processing of where a result came from and the specific datasets used in the workflow. In other words, the references reflect the lineage of the process.



Data citation examples
a) An article cites a dataset, b) a dataset is derived from two other datasets, c) subsets of a dataset are generated.

Within both citation approaches described above, there is variety in citable content type and granularity; for example, file level versus

dataset level. Repositories hosting these datasets can index these linkages and relationships within DataCite, as publishers do with Crossref.

These citation approaches reflect the how of data citations. The Force11 Joint Declaration of Data Citation Principles (JDDCP) connects the how to their purpose, function, and attributes. Two of the eight principles described in the JDCCP describe the value provided by data citation: *Credit and attribution* [1] (principle 2) and *Specificity and verifiability* (principle 7):

Joint Declaration of Data Citation Principles (excerpt)

Principle 2: Credit and attribution
Data citations should facilitate giving scholarly credit and normative and legal attribution to all contributors to the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.

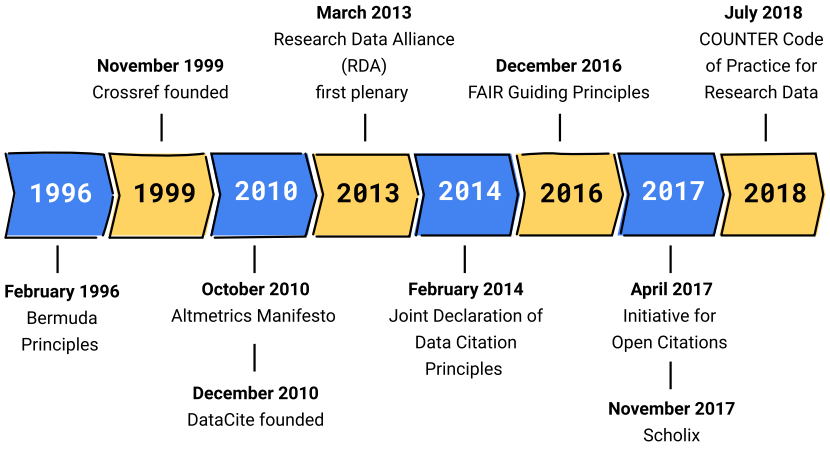
Principle 7: Specificity and verifiability
Data citations should facilitate identification of, access to, and verification of the specific data that support a claim. Citations or citation metadata should include information about provenance and fixity, sufficient to facilitate verifying that the specific timeslice, version and/or granular portion of data retrieved subsequently is the same as was originally cited.

Notes

1. While this book does not specifically address contributor roles in scholarly communications, it is understood that various author roles — e.g. data collection and data analysis — with regards to research data, require additional work. More information can be found at: Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211

Milestones and how we got here

Open data metrics are part of the larger open science movement and build on the work of several players in the research and scholarly infrastructure communities. While there is no way to capture all the moments that have enabled this work to progress, it is important to offer acknowledgement of key initiatives and organizations that have contributed to the development and understanding of data metrics. More detailed histories of the topic are outlined in Crosas (2014) [1] and Parsons (2019). [2]



Open data metrics milestones

The Bermuda Principles

In February 1996, the genome community met in Bermuda to formulate principles for circulating genomic data generated with funding from the Human Genome Project. The community agreed to release human genomic DNA sequences produced by large-scale DNA sequencing centers as rapidly as possible, and to submit finished data to the public sequence databases.[3] This outcome is an important milestone towards making research data available via a public database by default, and this now includes many more data types besides human genomic sequences.

Crossref

The non-profit membership organization Crossref[4] was founded in November 1999 to solve the problem of cross-referencing scholarly publications in the digital age. Crossref introduced Digital Object Identifiers (DOIs) as citation identifiers with required metadata, acting as an intermediary among publishers who have their publications reference each other. With the formation of this organization, community infrastructure for publishers to report the connections among research outputs became available.

DataCite

DataCite was founded in December 2009 to improve access to research data, facilitate data citation, and strengthen the importance of

research data as scholarly output. As a non-profit membership organization, DataCite builds functionality for data linkages and data usage that are complementary to Crossref, working together to provide open infrastructure. DataCite has invested in several data metrics-related initiatives — THOR,[5] FREYA[6] and Make Data Count[7] — three key projects focused on the linkages between scholarly works.

Altmetrics Manifesto

The Altmetrics Manifesto,[8] published in October 2010, provided a succinct description and roadmap for the then-new field of altmetrics. Altmetrics are bibliometric indicators complementing citations, usage, and peer review. They help with filtering the ever-increasing number of scholarly outputs in real time, and cover different facets of attention and impact, and for scholarly outputs beyond publications, including research data. The Altmetrics Manifesto initiated a number of tools and services, and a vast body of bibliometrics research, laying the foundation for how our communities record events associated with all research outputs.

Research Data Alliance

The Research Data Alliance (RDA) is a research community organization started in March 2013 by funders from Europe, North America, and Australia with the goal of building the social and technical infrastructure to enable open sharing and reuse of data. In 2014, the RDA/WDS Publishing Data Bibliometrics working group produced a survey on researchers' needs for data metrics,[9] establishing the Make Data Count project. In 2016, the RDA Data

Citation Working Group published recommendations for citing dynamic data.[10] In 2017, the RDA Data Usage Metrics working group began focusing on the adoption and development of data metrics.

Joint Declaration of Data Citation Principles

In 2012, representatives within the Earth Science Information Partners (ESIP) community proposed the need for data citation and published a set of data citation guidelines the same year.[11] In 2019, a group of authors versioned these guidelines to consider emerging use cases in data such as data versioning.[12]

The Committee on Data for Science and Technology (CODATA)[13] was established in 1966 as an interdisciplinary committee of the International Council for Science. In 2010, CODATA convened a joint task force with the International Council for Scientific and Technical Information (ICSTI)[14] to work on Data Citation Standards and Practices. In September 2013, the task force published its report.[15]

FORCE11[16] is a non-profit community organization of scholars, librarians, archivists, publishers and research funders founded in 2011 that arose organically to help facilitate the change toward improved knowledge creation and sharing. Building on the work by ESIP and CODATA, Force11 started a community initiative for a single set of data citation principles, and in March 2014, Force11 published the Joint Declaration of Data Citation Principles (JDDCP).[17] The 8 principles cover purpose, function, and attributes of citations, with the goal of encouraging communities to develop practices and tools that embody uniform data citation principles. As of October 31, 2019, the declaration has been endorsed by 120 organizations.

FAIR Guiding Principles for scientific data management and stewardship

Also arising from work within Force11, the FAIR Guiding Principles for scientific data management and stewardship were jointly developed by a diverse set of stakeholders and published in December 2016. [18] Their goal is to improve the infrastructure supporting the reuse of scholarly data, and they put specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals.

The FAIR guiding principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the datait describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Force11 Data Citation Implementation Pilot

The Force11 Data Citation Implementation Pilot (DCIP) was a community project started in February 2016, coordinated by Force11 members, and funded by a National Institutes of Health (NIH) grant, providing globally unique persistent identifiers for biomedical data., [19] and implementation guidelines of the Joint Declaration of Data Citation Principles for publishers [20] and data repositories [21]

Initiative for Open Citations

The Initiative for Open Citations (I4OC) [22] is a collaboration between scholarly publishers, researchers, and other interested parties to promote the unrestricted availability of scholarly citation data. I4OC launched in April 2017. As of September 2019, the percentage of publications with open references has grown from 1% to 59%.

Scholix

In November 2017, the RDA/WDS Scholarly Link Exchange Working Group (Scholix) [23] published a metadata schema describing data citations. Scholix has been implemented by a number of organizations, becoming the community standard for describing and exchanging data citation information.

COUNTER Code of Practice for Research Data

Make Data Count (MDC), an initiative between California Digital Library, DataCite, and DataONE, has focused on the standardization of data usage at repositories and promotion of proper data citation practices for publishers. [24] Through this project, the team partnered with COUNTER [25] — a non-profit organization developing standards for reporting the usage of scholarly resources — and published a Code of Practice for Research Data in July 2018 [26] This standard for defining data usage was coupled with a framework to report data views, downloads, and citations. Infrastructure providers DataCite and Crossref have jointly built a service called Event Data that provides an open hub for reporting citations and usage for datasets and publications respectively. The Make Data Count project leverages this service as a place to aggregate data usage counts.

Notes

1. Crosas, M. (2014). The Evolution of Data Citation: From Principles to Implementation. *IASSIST Quarterly*, *37*(1–4), 62. https://doi.org/10.29173/iq504

2. Parsons, M. A., Duerr, R. E., & Jones, M. B. (2019). The History and Future of Data Citation in Practice. *Data Science Journal*, 18(1), 52. https://doi.org/10.5334/dsj-2019-052.

3. Statement on the Rapid Release of Genomic DNA Sequence. (1998). *Genome Research*, *8*(5), 413–413. https://doi.org/10.1101/gr.8.5.413

4. Crossref [Website]. (2000). Retrieved November 2, 2019, from Crossref website: https://www.crossref.org/

5. Project THOR — Technical and Human infrastructure for Open Research. (2015). Retrieved November 1, 2019, from Project THOR website: https://project-thor.eu/

6. FREYA — Connected Open Identifiers for Discovery, Access and Use of Research Resources. (2018). Retrieved November 1, 2019, from https://www.project-freya.eu/en

7. Make Data Count. (2017). Retrieved November 1, 2019, from https://makedatacount.org/

8. Priem, Jason, T., Dario, Groth, Paul, Neylon, Cameron. (2010, October 26). Altmetrics: A manifesto — altmetrics.org. Retrieved October 29, 2019, from http://altmetrics.org/manifesto/

9. Kratz, J. E., & Strasser, C. (2015). Making data count. *Scientific Data*, *2*, 150039. https://doi.org/10.1038/sdata.2015.39

10. Rauber, A., Asmi, A., Uytvanck, D. V., & Proell, S. (2016). *Data Citation of Evolving Data: Recommendations of the RDA Working Group on Data Citation (WGDC)*. https://doi.org/10.15497/RDA00016

11. Data Stewardship Committee. (2012). Data Citation Guidelines for Data Providers and Archives. *ESIP*. https://doi.org/10.7269/P34F1NNJ

12. ESIP Data Preservation and Stewardship Committee. (2019). Data Citation Guidelines for Earth Science Data , Version 2. *Figshare*. https://doi.org/10.6084/M9.FIGSHARE.8441816

13. Committee on Data for Science and Technology — (http://www.codata.org/)

14. International Council for Scientific and Technical Information — (http://www.icsti.org/)

15. CODATA-ICSTI Task Group on Data Citation Standards and Practices. (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, *12*(0), CIDCR1–CIDCR75. https://doi.org/10.2481/dsj.OSOM13-043

16. FORCE11 — (https://www.force11.org/)

17. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 https://doi.org/10.25490/a97f-egyk

18. Wilkinson, M. D., Dumontier, M., Aalbersberg, IJ. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

19. Wimalaratne, S. M., Juty, N., Kunze, J., Janée, G., McMurry, J. A., Beard, N., … Clark, T. (2018). Uniform resolution of compact identifiers for biomedical data. *Scientific Data*, *5*(1), 180029. https://doi.org/10.1038/sdata.2018.29

20. Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., … Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*(1), 180259. https://doi.org/10.1038/sdata.2018.259

21. Fenner, M., Crosas, M., Grethe, J. S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., … Clark, T. (2019). A data citation roadmap for scholarly data repositories. *Scientific Data*, *6*(1). https://doi.org/10.1038/s41597-019-0031-8

22. Initiative for Open Citations (I4OC). (2017). Retrieved November 1, 2019, from https://i4oc.org/

23. Burton, A., Fenner, M., Haak, W., & Manghi, P. (2017). Scholix Metadata Schema for Exchange of Scholarly Communication Links. *Zenodo*. https://doi.org/10.5281/ZENODO.1120265

24. Make Data Count. (2017). Retrieved November 1, 2019, from https://makedatacount.org/

25. COUNTER — Consistent, Credible, Comparable. (2019). Retrieved November 1, 2019, from Project Counter website: https://www.projectcounter.org/

26. Fenner, M., Lowenberg, D., Jones, M. B., Needham, P., Vieglais, D., Abrams, S., … Chodacki, J. (2018). The COUNTER Code of Practice for Research Data. *Project Counter*. Retrieved from https://www.projectcounter.org/code-of-practice-rd-sections/foreword/

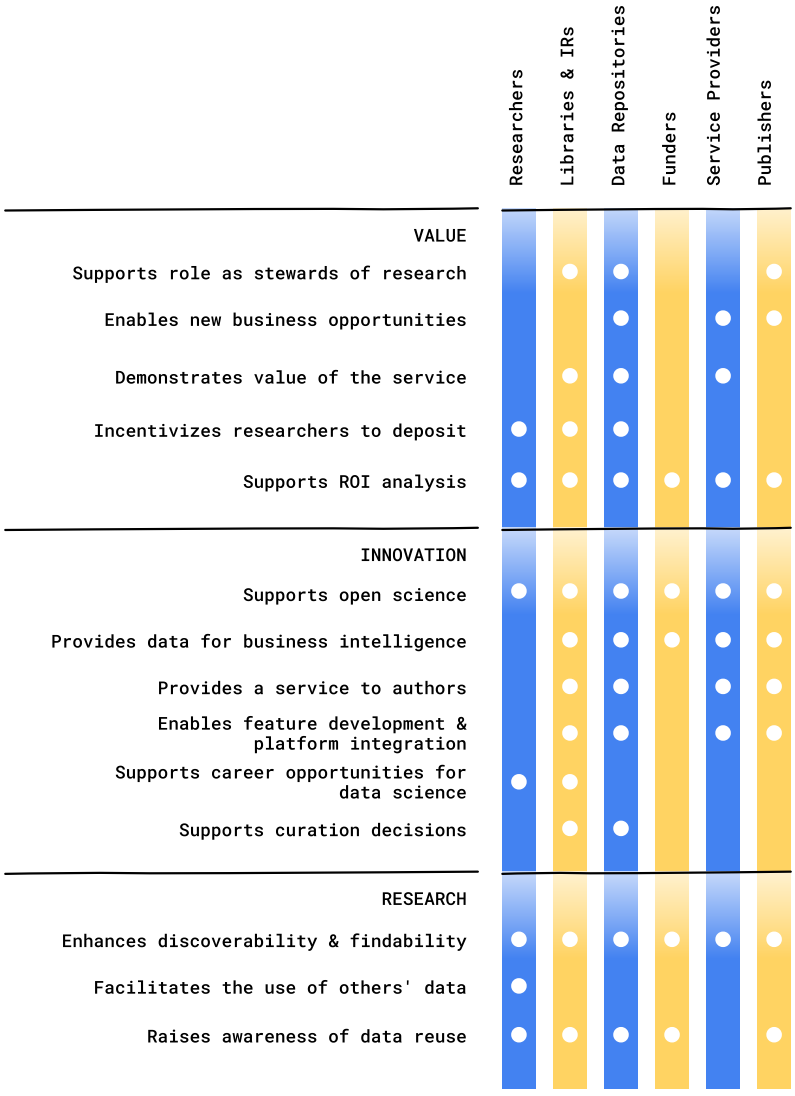
Value to the community

Research data is important in driving advances in science, policy, and management. By virtue of this, there is a need for an objective basis on which to evaluate investments made in research. To date, the research community has not had a consistent, properly contextualized understanding of the impact and utility of research data. This is largely because the community has not agreed upon common and widely applicable frameworks for data metrics. However, it is possible to get to a point where open, trusted, and well understood data metrics are broadly adopted and become the norm.

Embarking on the journey to open data metrics, it's important to consider the value that community-built and -adopted data metrics can bring. The benefits for different players may vary, but there are key values of data metrics that are shared across communities. These include:

- Enhanced discoverability and findability
- A clear understanding of the impact of shared data
- Further research into the science of science
- Better data for business intelligence around research
- Business opportunities and services leveraging this shared data and knowledge

In addition to these, we can also foresee additional benefits that pertain to specific stakeholder communities.



Benefits by stakeholder

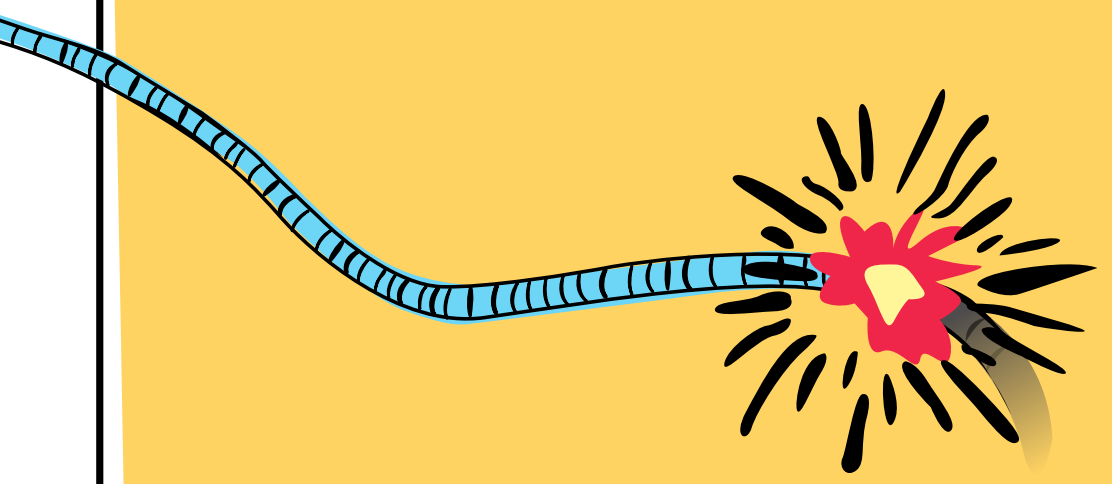
Sample visualization of open data metrics benefits by stakeholder group

Each organization in the research community wants to understand more about their constituents. Having data, both quantitative and qualitative understanding of behaviors, would help us to understand how use of data would be welcomed by organizations, commercial and otherwise. Providing robust services is easier with a better understanding of users and their needs.

As data sharing is increasingly common, understanding the scope and impact of data reuse and citation are crucial. Funders, for example, have particular interest in tracking the sharing of data supported by their grants. Other stakeholders have an interest in analyzing community networks and practices within disciplines and possibly smaller subcommunities. On a more technical level, others will want to analyze versions of shared, distributed data that could illuminate how to improve infrastructure, hosting, and discoverability considerations. Taking data repositories as another example, metrics could enhance their platforms, which could offer additional or more robust services to researchers and institutions.

The research community can also look toward this space for creative uses of metrics and incentives, should they benefit their disciplines and research goals. The aspiration of having non-traditional outputs like research data included in the tenure and promotion process could be enabled by having a trusted data metrics system. Certainly there are disciplinary differences that should be acknowledged and broader representation from across all subjects, but researchers across the board can benefit from the development and adoption of open data metrics. Similarly, input from groups such as funders, publishers, and researchers is needed to identify goals and benefits of having a maturely developed data metrics ecosystem. Further engagement from all groups can help contribute to both the case for, and development of, data metrics with future benefits in mind.

3 Coordinating emerging tools and standards



To realize the benefits of metrics in helping us to understand the impact of research data, there needs to be practical implementation of tools and services for those metrics. While there is no single solution that will accommodate all possible use cases for research data, a mature ecosystem of services is evolving. The biggest gains will come from community-developed standards, tools, and services guided by a shared direction and vision for a future of open data metrics. Approaches towards this future state must remain researcher-focused, have easily understandable motivations, and be easily implementable. This section describes ideals for standardized and transparent data citation and data usage services, the ecosystem of services that has already been developed and tested for data metrics, and the adoption and deployment of this data metrics infrastructure in various communities.

Ideals of our framework

Moving a community to adopt a particular set of ideals is impossible without a clear articulation of the value of those ideals. For data metrics, many people debate whether the resulting system will be worth the implementation costs. While these discussions need to be fleshed out across participating communities, below are proposed ideals on which data metrics standards and services could be based:

1. Open data metrics are *open*. They can be freely used, shared, and built-on by anyone, anywhere, for any purpose. They are made available under an appropriate open license to make this explicit.

2. Open data metrics are *inclusive*. They can be generated, reported and used in all disciplines, geographic areas, and communities, and for all data types.

3. Open data metrics are *structured*. They are expressed in machine-readable formats and can be accessed programmatically.

4. Open data metrics are *transparent*. They are based on open standards, and their generation and reporting are documented openly to the full extent made possible by laws and ethics.

5. Open data metrics are *interoperable*. They are open, structured, and transparent, allowing for the aggregation of data from multiple sources.

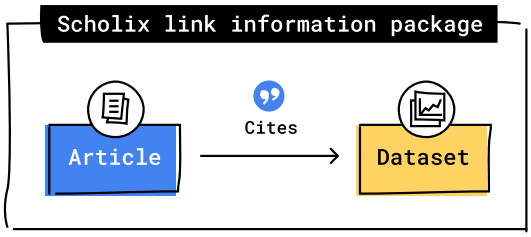
6. Open data metrics are *multi-dimensional*. They reflect the multitude of ways research data can be reused, and don't conflate these dimensions into a single metric or number.

Implementation of an open framework

There has been considerable community progress on the information standards for representing citation and usage data, the aggregation services for compiling and indexing metrics, and the access services making the compiled metrics available to the community. In general, it is critical for interoperability and comparability that services follow these standards. For data citations, the Scholarly Link Exchange (Scholix) [1] has been the focus of community agreement. In addition, for data usage, the COUNTER Code of Practice for Research Data [2] has been the emphasis of community discussion and implementation.

Data citation tracking

The fundamental challenge with data citation is the large number of disconnected data sources that hold citation information (publishers, repositories, infrastructure providers, etc.) and the heterogeneity of citation practices (different ways of referencing data, different persistent identifier systems, different events when data were cited, etc.). To overcome these challenges, a large group of organizations from the data repository and publisher communities came together via an RDA working group to develop a standard framework for reporting data citations: The Scholarly Link Exchange (Scholix) metadata schema defines how data citations should be reported.



A Scholix Link Information Package
The package contains information about the two objects, and information about the nature of the link and the link package itself

Scholix Link Information Packages represent the core set of relationships that can be used to reference a link between an article and a dataset, or between two articles, or two datasets, etc. For example, the Scholix Link Information Package image above, shows a linked relationship in which an article 'cites' a dataset, but other relationships can be recorded as well. By standardizing the way these links are expressed, Scholix allows the highly diverse providers of this information to report their link data to Scholix Hubs, which in turn share the links with other hubs, and with Scholix consumers such as repositories, publishers, and service providers. The two Scholix Hubs, as of October 2019, are the ScholeXplorer (OpenAIRE)[3] and the Crossref/DataCite Event Data[4] service, each of which collates, aggregates, and reports on these link relationships for the broader community. These two hubs both follow the Scholix standard, and both exchange Scholix Link Information Packages between each other, but also make them available via open APIs.

Data citation information in the Crossref/DataCite Event Data service comes from DOI metadata for DOIs registered by publishers and data repositories, both from Crossref and DataCite. Thus, data citations in a

journal article are reported via Crossref, and data citations provided by a data repository are reported via DataCite.

With Crossref DOIs, information about data citations can be included with references and relations metadata, the latter allowing the use of a relation type describing the relationship between publication and data. With DataCite DOIs, information about data citations can be included with *relatedIdentifier* metadata, again allowing the use of a relation type describing the relationship between publication and data.

Data usage tracking

Consistently reporting on data usage is complicated considering the variety of ways to define data and the variety of practices for recording usage. To overcome these challenges, the COUNTER Code of Practice for Research Data was created to standardize the generation and distribution of usage counts for research data, enabling, for the first time, consistent and credible reporting of research data usage.

These counts are normalized representations of the number of times that datasets have been viewed and downloaded, accounting for differences in practice among data providers. The COUNTER Code of Practice for Research Data provides the guidance needed to:

- Standardize logging usage events
- Log processing to extract meaningful counts
- Report usage data

The code of practice is aligned as much as possible with the COUNTER Code of Practice for Release [5] for standardized reporting of publication usage metrics. The Code of Practice for Research Data, which was released in July 2018, has already been adopted by a number of data repositories, repository platforms and aggregators, including Dryad, Zenodo, Dataverse, DataONE, and Caltech.

Log processing for usage metrics is not easy and there is more work needed for shared community approaches and tools that promote the adoption of the Code of Practice for Research Data by repositories. DataCite provides a service for collating the Code of Practice for Research Data Usage Metrics reports in a centralized hub. These reports are in turn made available for download, but also processed into a format aligned with the Scholix format and made available via the Crossref/DataCite Event Data service.

By aggregating usage reports across providers, data owners and interested researchers can gain a more complete picture of the views and downloads that have occurred for a given dataset. This central corpus of aggregated data usage counts can be utilized by a variety of stakeholders, following the above ideals. Essential to all this is that this hub holds standardized and transparent usage reports, creating a corpus that researchers and institutions can trust.

As the hub of this large and diverse pool of data usage and citation counts, Event Data allows for bibliometricians, data scientists, and those interested in research on trends in data, to analyze and build useful results that span the corpus. With this information, infrastructure and services can be built that provide those results back to their user communities. For example, Event Data is especially useful to aggregators and repositories as it provides access to collated usage and citation counts for datasets that are viewable or downloadable

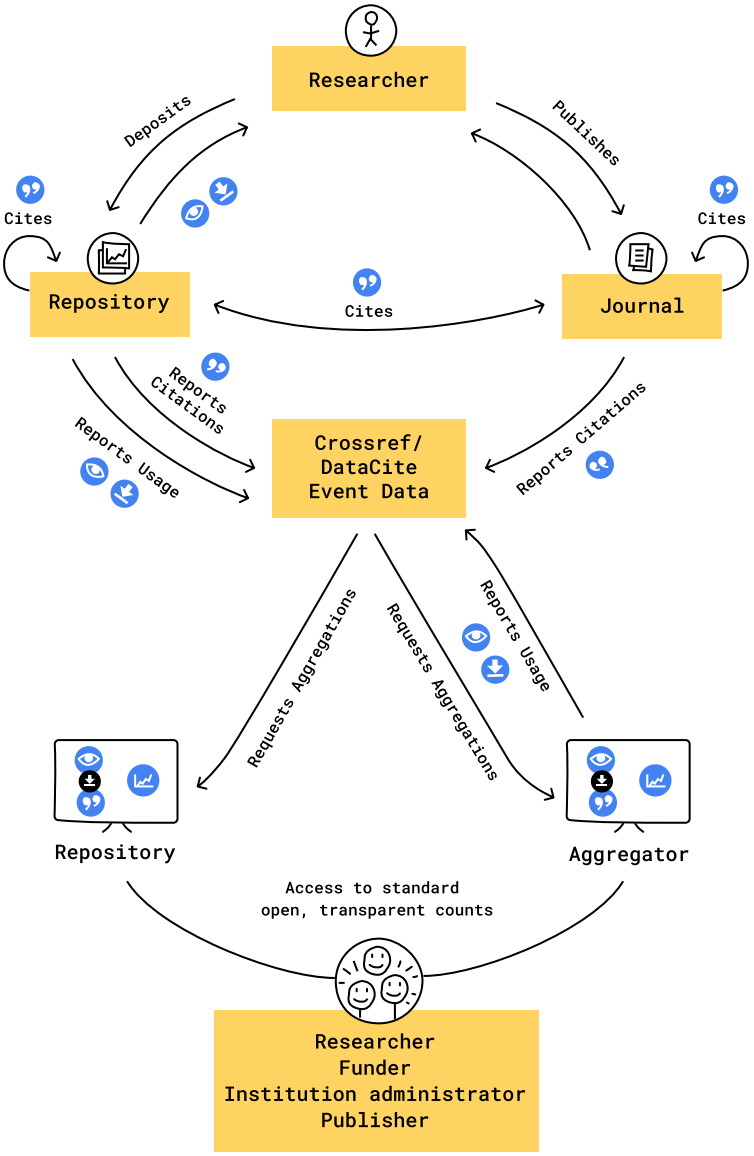
from multiple locations (e.g., from the authoritative repository, and from an aggregated search service). As further repositories contribute to this corpus, it will also grow more useful for combined reports across funding agencies, data creators and authors, and repositories.

Bringing it all together

There are various ways these standards can be put into action. The figure below captures the above standards, illustrating how implementation of this emerging framework can be useful to researchers, repositories, publishers, and other stakeholders. At the top of the figure, researchers create new work, depositing datasets into various data repositories, and publishing articles and other scholarly works in various journals. These works are the core of the research data metrics ecosystem.

Whenever other researchers access datasets from repositories, repositories can record those views and downloads following the standards defined in the COUNTER Code of Practice for Research Data, and then periodically send these standardized usage reports to DataCite for indexing. Likewise, data aggregators that provide network-wide search and discovery can also report views and accesses to any replicas of the data and metadata that they hold, and report these back to DataCite following the standard report format.

In parallel, when researchers use a dataset within their work, they can cite the dataset in their article. The associated journal would then record that citation in the metadata associated with its DOI and send that to Crossref when they register the DOI for the article. These reported citations are then extracted following the Scholix standards, and exchanged among the various Scholix Hubs, making the link information available to the broader community.



Framework for standardized data usage and citation

Combining standards (Scholix, COUNTER Code of Practice for Research Data) and open infrastructure to report and display normalized and aggregated research data usage and citations.

In this way, both citation and usage data are collated and aggregated, and become available through Event Data and similar services. This standardized data citation and data usage information becomes valuable to all stakeholders for displaying the importance and connections of research data across the community. For example, repositories query the Event Data service and provide data usage summaries on their dataset landing pages, as well as lists and links to articles that cited the data. Journals provide links from their article pages back to the data that they cited, both directly and indirectly by understanding the citation graph and crediting datasets deep from a researcher's workflow that were critical to the findings in the article. Data aggregators and search services could likely provide these same types of services, but also provide aggregated reports that show cumulative usage and citation trends over time to inform researchers, funders, administrators, and publishers about the changing impact of data holding on research. Finally, DataCite itself displays data citations and data usage in its search service, making use of the information stored in the Crossref/DataCite Event Data service.

As of October 31, 2019, the Crossref/DataCite Event Data service had captured 1,290,962 unique dataset views and 239,079 unique dataset downloads from 24 repositories, and 2,460,788 data citations, the vast majority reported by data repositories, and only 7,589 data citations reported by publishers. The two most widely used relation types are references and *isSupplementTo*. These numbers demonstrate both that the Crossref/DataCite Event Data service is already capturing and reporting significant numbers of data usage and data citations, but also that there is more adoption work needed, both in the number of repositories reporting data usage, and in the number of data citations reported by publishers.

Notes

1. Burton, A., Fenner, M., Haak, W., & Manghi, P. (2017). Scholix Metadata Schema for Exchange of Scholarly Communication Links. *Zenodo*. https://doi.org/10.5281/ZENODO.1120265

2. Fenner, M., Lowenberg, D., Jones, M. B., Needham, P., Vieglais, D., Abrams, S., … Chodacki, J. (2018). The COUNTER Code of Practice for Research Data. *Project Counter*. Retrieved from https://www.projectcounter.org/code-of-practice-rd-sections/foreword/

3. ScholeXplorer — The Data Literature Interlinking Service. (2019). Retrieved November 1, 2019, from http://scholexplorer.openaire.eu/#/

4. Event Data — Open for your interpretation [Website]. (2019). Retrieved November 1, 2019, from Crossref website: https://www.crossref.org/services/event-data/

5. Fenner, M., Lowenberg, D., Jones, M. B., Needham, P., Vieglais, D., Abrams, S., … Chodacki, J. (2018). The COUNTER Code of Practice for Research Data. *Project Counter*. Retrieved from https://www.projectcounter.org/code-of-practice-rd-sections/foreword/

Pioneering efforts

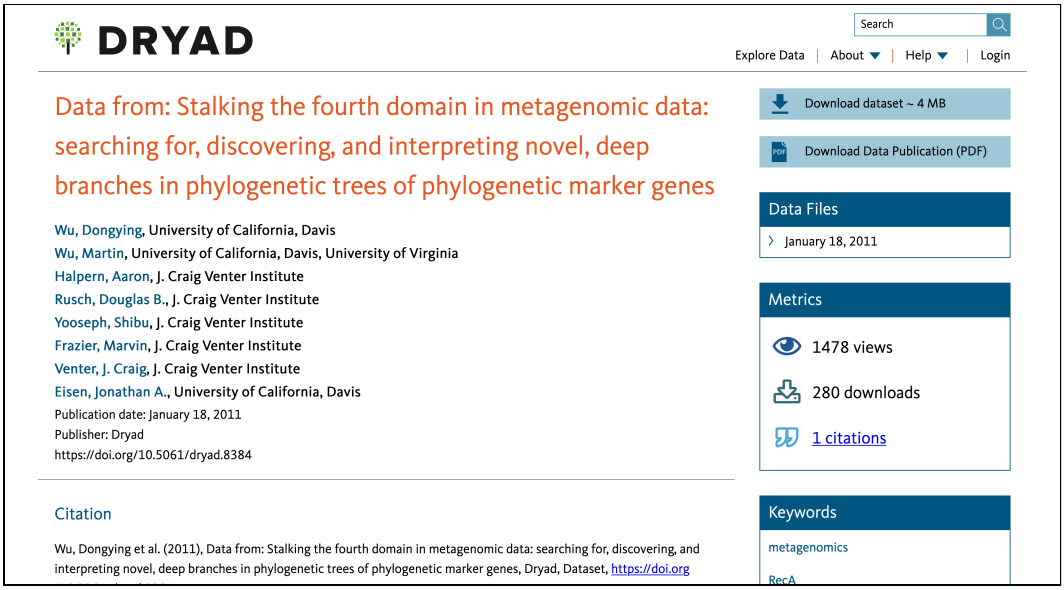
For standards such as Scholix and the COUNTER Code of Practice for Research Data to become solutions, they need to be widely adopted. A critical mass of available data citations and usage counts, regular reporting of these counts, and studies on correlations between and within data usage and citation are necessary for systems and services in scholarly communications to rely on this information. In other words, these components need to be, by default, adopted by the community.

There is broad agreement on the beneficial values of getting to a state where we can have research data metrics. This consensus needs to be matched by participation. Of course, adoption practices vary by stakeholder. This means researchers need to be thinking about citing, referencing, and linking datasets to their other research outputs. Repositories hosting these datasets should index these linkages, standardize their view and download counts, report this information to an open hub, and display this information back for researchers and others on dataset landing pages. Publishers and repositories can encourage data citation both as promoting citation as an accepted practice, and by indexing their article-data relationships with Crossref.

The framework presented in the last section has been implemented in standalone repositories as well as repository networks. Spotlighting a couple of these implementations shows the value of standardizing and reporting data usage and citation.

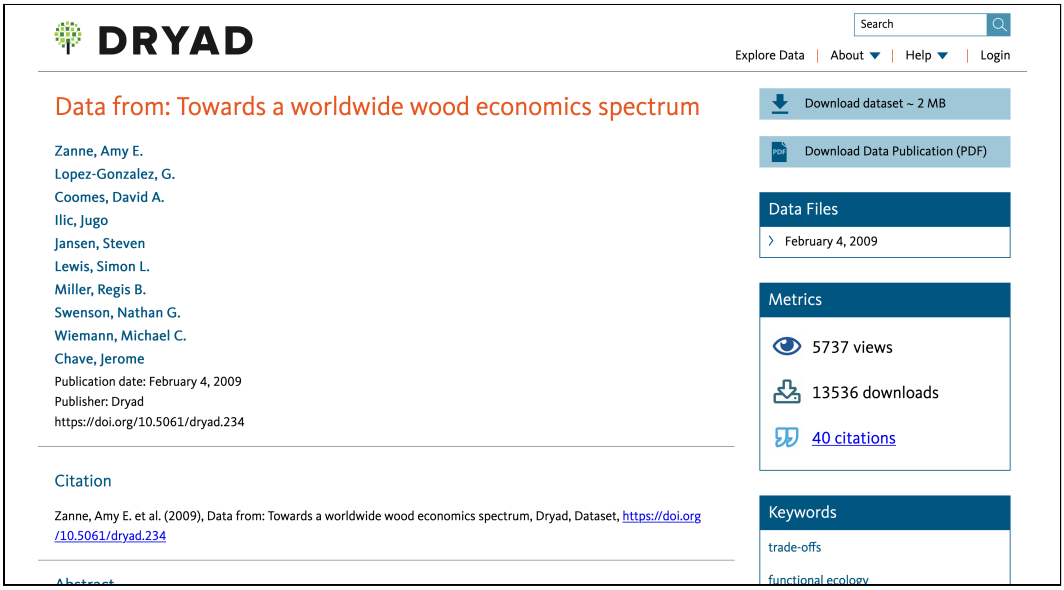
Dryad

Dryad, a curated data repository, reviewed ten years of log files to standardize their usage with the COUNTER Code of Practice for Research Data. Sending these normalized files to DataCite has now enabled aggregators and those interested in data statistics (i.e. bibliometricians) to access a large corpus (30,000 datasets) of usage information. Displaying these views and downloads has also assured researchers that their counts are not inflated by bots or other agents. Utilizing this open hub at Event Data, Dryad is now able to display — back to researchers — the citations of their datasets, regardless of when their dataset was published. By completing the cycle of normalizing counts, reporting them to DataCite, and displaying the aggregated statistics, stakeholders such as researchers, institutions, publishers, and other supporters of the research process can now have access to this information in a trusted and open manner.



Example Dryad Dataset 1

Displaying the result of normalized views and downloads on a dataset at Dryad



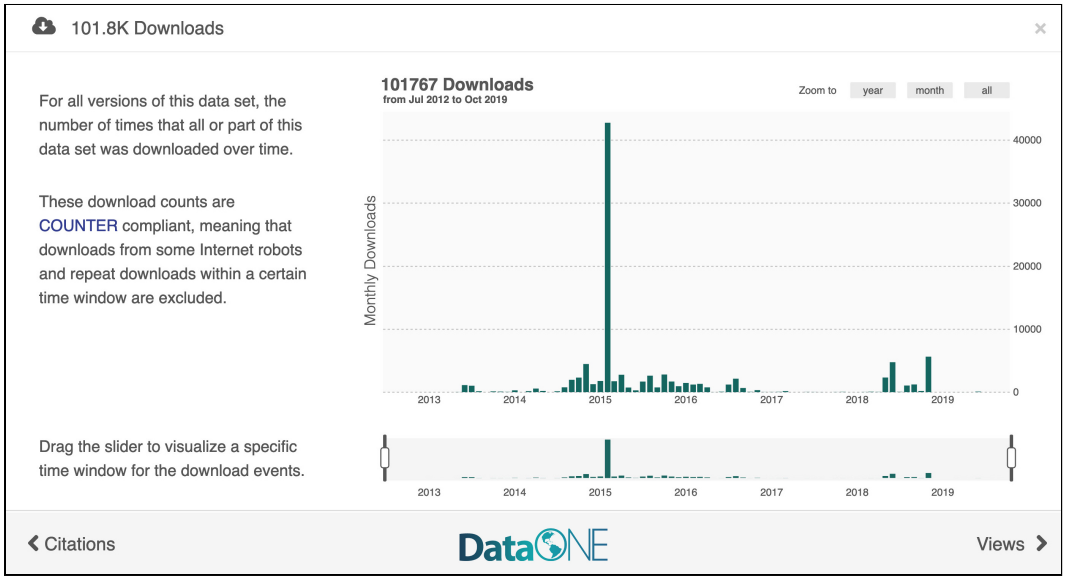
Example Dryad Dataset 2

Exposing the reach of datasets by pulling in citations from Event Data

DataONE

DataONE is a network of repositories designed to provide shared data discovery and access from a data portal that appears centralized but which in fact provides access to data distributed throughout the DataONE network. Users can view and directly access datasets at repositories, or through replicated copies of the housed data at aggregators such as DataONE. Since data and metadata are replicated to DataONE and other aggregators, without the Make Data Count framework, the authoritative repository might not be aware of all data views and downloads. At this network level, DataONE has shown how to responsibly replicate data by reporting views and accesses to the open Event Data portal.

Aggregating these repository-reported usage counts with the usage data reported from DataONE ensures that datasets in repositories involved in networks get a complete picture of their usage. In addition, DataONE harvests access logs from a portion of their forty-three member nodes and provides a service to normalize these and report usage to DataCite on behalf of the member repositories, thereby streamlining the process for members. For repositories that contribute to the DataONE network, they can rely on the fact that DataONE reports these numbers to Event Data for dataset level, repository level, and other levels of aggregation.



Example DataONE dataset usage visualization

Displays the standardized download counts over time showing unusual spikes, in this case associated with a large machine learning class using the same large dataset

Addressing implementation challenges

While it possible to showcase the benefits of participating in the emerging open framework for data usage and citation, adoption is not widespread, and it is important to acknowledge the barriers in order to mitigate them.

Repository implementation

Implementation requires time and resources that need to fit with competing priorities. Communities such as the RDA Data Usage Metrics Working Group [1] have reported that cost and prioritization are barriers to implementation. Some repositories have implemented pieces of the emerging open framework but not others. To increase adoption, barriers need to be lowered and data metrics collection made easier to implement, e.g. by improving open source tools and documentation for log processing and usage reporting, and by offering data usage reporting as a service.

Repositories can host both datasets and publications in the same repository. While the repositories may understand there are differences in the way that researchers will cite and use the different outputs, they

may use a single approach to track metrics. The workflow for tracking citations is very similar, but for tracking and reporting the usage of datasets and articles we need to utilize different COUNTER Codes of Practice.

Implementation across disciplines

Another barrier to adoption relates to the use of specialized data types across disciplines. Different levels of granularity across datasets are prevalent within and across these disciplines. For example, because genomics researchers may use and cite individual nucleotide sequences, while earth science researchers use and cite expansive datasets that encompass billions of observations, there are natural differences of interpretation in what an individual download or citation means. Downloading a single sequence is not equivalent to downloading the whole human genome and citing a museum record is not equivalent to citing a dataset detailing global biodiversity patterns. Despite these differences in interpretation, disciplines would still benefit from standardizing how search engine robots are filtered from usage data, and how download sessions are handled, when tracking usage.

Similarly, there are disciplinary differences in the use of persistent identifiers for data. Many disciplines have converged on using DOIs to identify data but in the life sciences, compact identifiers [2] are much more widely used. The type of identifier is not relevant when tracking usage data but requires different approaches for tracking data citation. The proposed open framework accommodates these disciplinary differences; for example, the ScholeXplorer [3] service tracks data citations using compact identifiers.

Beyond technical barriers like the persistent identifier used for datasets, this framework does not address datasets that are sensitive, purchased, or are accessed through a mediated service (e.g. human subject research). In these circumstances, where the ability to download the dataset is not available to the public, we cannot responsibly account for these data usage counts. Engagement with repositories that host these types of content could help to define these use cases and associated behaviors so that they can be included in the current standards and framework.

While some disciplinary repositories may deprioritize adoption due to these difficulties of interpretation or implementation, the community benefits when usage metrics are standardized as much as possible and regardless of these differences, should continue to strive towards the core ideals for standard, open, transparent, and accessible metrics across all disciplines.

Publisher implementation and framework support

Publishers are uniquely positioned to support their authors and all of scholarly communications by facilitating data citation. In the last decade, there has been widespread acknowledgement that prominent data availability statements on journal articles are a priority.[4] The implementation of these data availability statements, as a result of journal data policies, has increased the awareness of article-related datasets.[5] Author reluctance to comply with journal data policies has continually decreased and feedback to the author community about their citations can further drive support for open data at the time of article publishing. Increasing the rates of publishers implementing machine-readable data statements and indexing these relationships as data citations will help to provide a fuller ecosystem of data citations, in turn providing publishers with a more comprehensive view of their published research and its reach.

Publishers still have work to do: many journals publish supplemental materials that are often the datasets underlying articles. These datasets should instead be deposited in public repositories that support FAIR discoverability, dataset-specific metadata, and long-term archiving. Publishers can guide authors towards appropriate data repositories in their author guidelines.[6] Also, far too often, data referenced in a publication are not included in the metadata for the publication sent to Crossref.

The role that all stakeholders can play in this adoption campaign is to engage on the topic and collaborate on open frameworks to support the common goal of assessing the reach and value of research data. Without choosing closed system approaches, all interested parties can weigh in on the barriers to adopting these practices so the community can work to diminish the barriers. If this active participation from a diverse pool of stakeholders is neglected, dataset reach is undermined and impact cannot be compared or understood.

Notes

1. RDA Data Usage Metrics WG. (2017, November 22). Retrieved November 1, 2019, from RDA website: https://www.rd-alliance.org/groups/data-usage-metrics-wg

2. McMurry, J. A., Juty, N., Blomberg, N., Burdett, T., Conlin, T., Conte, N., … Parkinson, H. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLOS Biology*, *15*(6), e2001414. https://doi.org/10.1371/journal.pbio.2001414

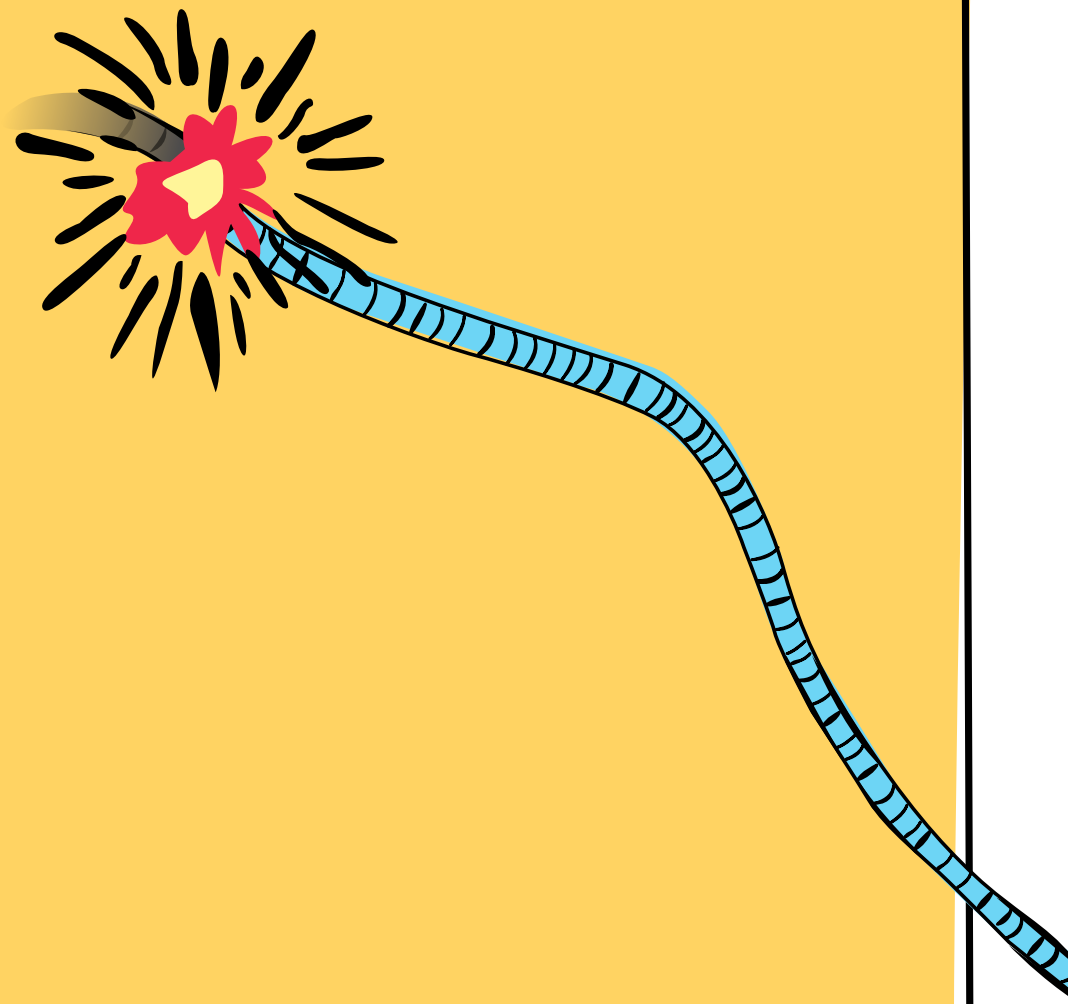
3. ScholeXplorer — The Data Literature Interlinking Service. (2019). Retrieved November 1, 2019, from http://scholexplorer.openaire.eu/#/

4. Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., … Clark, T. (2018). A data citation roadmap for scientific publishers. *Scientific Data*, *5*(1), 180259. https://doi.org/10.1038/sdata.2018.259

5. Federer, L. M., Belter, C. W., Joubert, D. J., Livinski, A., Lu, Y.-L., Snyders, L. N., & Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLOS ONE*, *13*(5), e0194768. https://doi.org/10.1371/journal.pone.0194768

6. Cousijn, H., McQuilton, P., & Sansone, S. (2019). MoU between DataCite and FAIRsharing: Improving criteria for the recommendation of data repositories. *DataCite Blog*. https://doi.org/10.5438/Z32P-WJ46

4 Avoiding traps

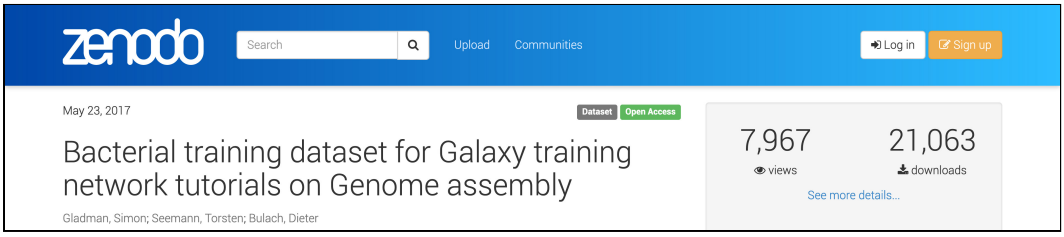


As with all research and scholarly communications spaces, the mechanisms that support and develop data metrics are evolving quickly. However, one unique characteristic of data metrics is that, as an emerging field, it can be developed in a responsible and considerate way. Looking at a couple of scenarios, however, there are potential pitfalls and situations that may play out in ways that do not benefit the community and may in fact disservice researchers.

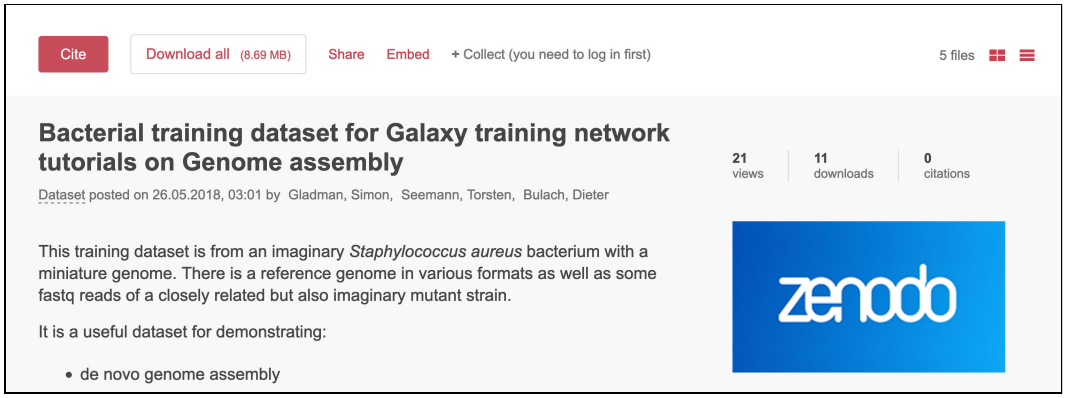
Aggregating responsibly

Aggregators are networks of repositories as well as data repositories that harvest datasets and their associated metadata. Regarding data metrics, aggregators collect views, downloads, and even citations to the data. Their involvement in metrics reporting is essential to building an accurate data metrics ecosystem.

These services mirror datasets across multiple locations and play a significant role in the data community. Since researchers view and download data from aggregator sites in addition to accessing data at the repository where it was originally published, they are key to improving data access. However, this only works as long as these views and downloads are, in return, normalized and reported back to the community. For completeness, these reported counts need to be available for the original repository to access, aggregate, and display (and vice versa) in order to ensure that researchers and other stakeholders gain a complete picture of their data usage.



Dataset originally published in Zenodo
Views and downloads are standardized against the COUNTER Code of Practice for Research Data and displayed



Same dataset harvested and mirrored at figshare
Displayed views and downloads that have not been normalized and do not include the comprehensive usage of the dataset, seen in the other figure

Evidenced in the two figures above, we see the potential for misleading researchers about dataset usage when an aggregator fails to report their usage back to a shared service such as Event Data, or only displays their own usage (e.g., 17 downloads) rather than the aggregated downloads including the authoritative repository's usage (e.g., > 20K downloads). The result is downstream services (i.e. business intelligence tools, etc.) that report only from a single source to paint an incomplete picture of data usage and citation. This example can happen in the opposite direction too, where aggregation systems receive higher volumes of views and downloads on a specific dataset but are not normalizing or reporting this usage.

In the above scenario, the original repository (where the citation points to) and others interested in these are left in the dark. In addition, the same dataset might also be available from multiple mirrored locations, something that is rather common for major life sciences databases such as the Protein Data Bank (PDB). [1] Only by collecting data usage and data citation in a standardized way, followed by aggregation of this information into a shared, open metrics hub, can the true extent of data reuse be understood.

The importance of proper reporting also pertains to institutional repositories where copies of research outputs published elsewhere are held. By ingesting content into their repositories without broadcasting new identifiers in order to avoid duplication of citations, institutional repositories often display the originally published identifier for the mirrored work they are hosting. This makes sense to support de-duplication of citations in scholarly works, but without reporting and displaying standardized usage counts, institutional repositories can fall into the same issues that we see above, where true usage and citation are not normalized, reported, and displayed.

Notes

1. wwPDB consortium, Burley, S. K., Berman, H. M., Bhikadiya, C., Bi, C., Chen, L., … Ioannidis, Y. E. (2019). Protein Data Bank: The single global archive for 3D macromolecular structure data. *Nucleic Acids Research*, *47*(D1), D520–D528. https://doi.org/10.1093/nar/gky949

Navigating the hype

In recent years, our communities have become inundated with products and services that promise easy solutions to complex problems. These projects have come from many corners of scholarly communications, including emerging data science entities. As with any industry, the reasons for this are multi-faceted. Even with shared goals, these services could have the potential to overpromise and mislead.

As outlined earlier in this book, data metrics are on a path towards development but are not yet mature enough to be used for strategic planning or impact assessment. However, there are many reasons why products and services could be positioned otherwise:

1. **Market pressures**. The marketplace is hungry for information on data outputs. Research offices, libraries, publishers, and funders are increasingly looking to understand the reach of their investments and/or the reach of their research. They turn to products that fill the void with promises of uniquely formulated business intelligence or the illusion of universally comparable metrics.

2. **Innovation pressures**. All sides of the scholarly communications ecosystem yearn to build clever, new widgets and nothing is more in vogue than data. With shifts in the business models of publishing, players are scrambling to find new ways to monetize their expertise and infrastructure. Publishers and libraries look to move upstream to the world of data as a way of getting closer to the research.

3. **Information gap**. Universities, funders, and researchers are being squeezed by taxpayers and other constituents to prove the value of research. Previous methods of measuring the impact of projects do not fully describe the realized and potential impact, and so the community is looking to the exploitation of data metrics to fill that void.

4. **Expertise gap**. Many organizations feel overwhelmed when it comes to working with research data. The unique requirements, described in the "Understanding data and data metrics" chapter about the ways to define a dataset and data usage, can lead to a consensus in the scholarly communications community that are not equipped to adapt and handle research data, especially considering this relatively fast-paced shift in industry priorities.

Current tools in the market typically do not include research data at all, or if they do, they do not take into consideration the complexities of granularity, data organization, and data derivation. Failing to do so can cause misinterpretations of research data impact. Similarly, libraries that are looking to build data curation departments should not assume that these pre-emptively defined metrics can be the basis for their decision-making. Researchers that are hoping to find a new, easy-to-understand roll-up number or index score must be careful not to base the value of research data on metrics that are themselves based on incomplete development.

While the industry shifts are exciting, and they bring with them innovation and renewed support for research data, providers intending to sell us incomplete projects as finished products, should be regarded with a healthy dose of skepticism. The research community should remain grounded and informed of the reality of the need for rigorous

and transparent data metrics. Research offices at institutions that are looking to compare data outputs from their universities should not assume that the business intelligence tools they just purchased are based on complete and/or normalized information.

There is a path forward for data metrics but it will take time and resources to achieve the vision. While current products are appealing, can alleviate the sense of urgency, and provide a quick interpretation of data usage and impact, the entire community is wise to remain cautious and to keep in mind the considerable work needed to achieve the ideals of an open and accessible data metrics ecosystem.

Being mindful of gaming

Any time new metrics are created, there are opportunities for gaming in the form of misuse and misinterpretation. [1] Creating data metrics, in a high stakes and broken scholarly reward system, will give rise to new gaming opportunities for researchers and supporters. It is important that this behavior not be incentivized.

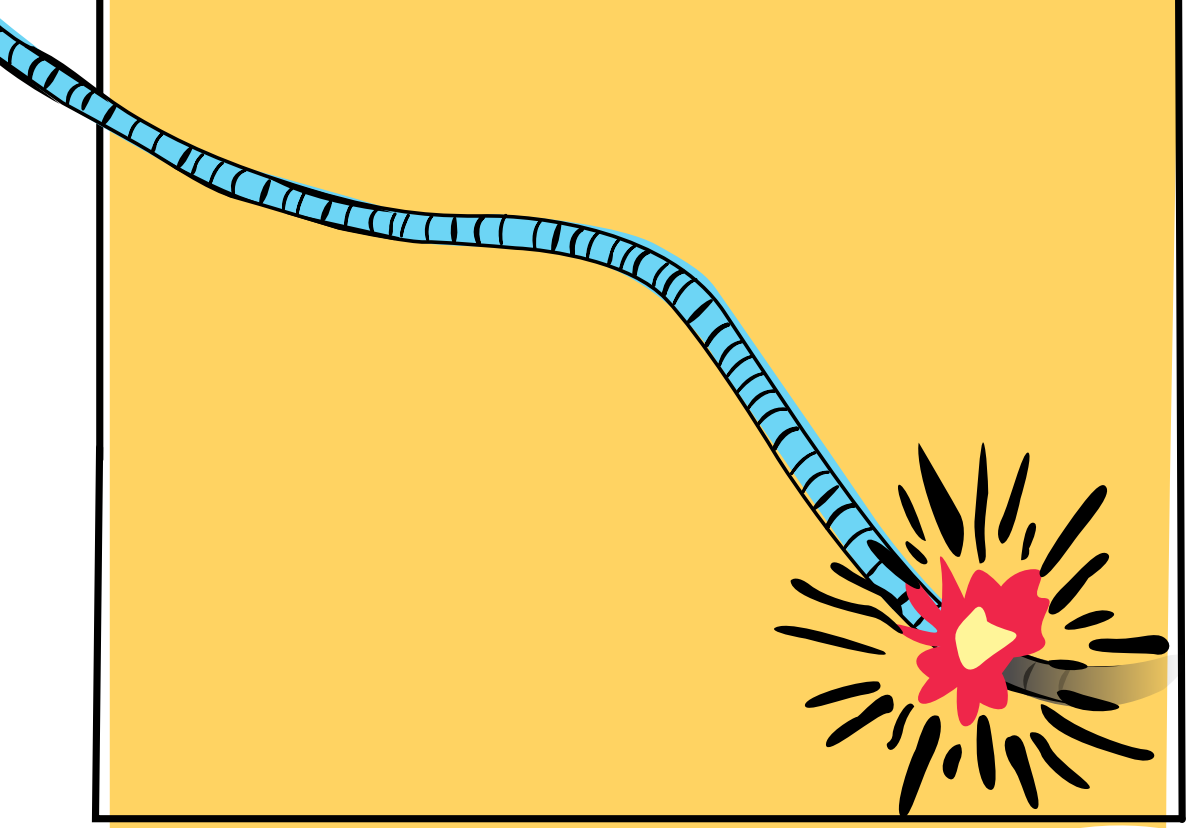
Metrics gaming can happen in a variety of ways. Gaming of publication metrics is a known problem and it would be naive to believe this won't happen with research data. However, as with other research outputs, there is no magic solution that will outright prevent this behavior. Humans and computers will always find ways around preventive measures. Instead, efforts can be focused on building a robust data metrics ecosystem that includes defensive approaches to gaming through community outreach, risk mitigation, and transparency.

The potential for gaming can be positioned as an insurmountable obstacle that should stop data metrics from moving forward. Instead of taking this approach, the community can work to avoid blatant traps and work to responsibly create metrics. While taken seriously, there can be a balance between both community agreement that gaming is not a supported behavior and with concentrated efforts toward creating data metrics systems that reward reuse and recognition of research data. This is feasible as long as the underlying information for data metrics are open, auditable, and transparent. Moreover, as these metrics develop, the community should work to remain proactive and agile to mitigate emerging risks and adjust to new approaches to gaming.

Notes

1. Gordon, G., Lin, J., Cave, R., & Dandrea, R. (2015). The Question of Data Integrity in Article-Level Metrics. *PLOS Biology*, *13*(8), e1002161. https://doi.org/10.1371/journal.pbio.1002161

5 The future of data metrics is bright



Research data is at the center of science, and to date it has been difficult to understand its impact. Properly valuing research data means building tools and services that make both sharing and reuse of research data easier. It also means incentivizing researchers to share and reuse research data, recognizing the complex and realistic research process, and giving attribution to those involved in data creation and analysis.

The community has already come a long way: our communities support and recognize the importance of data sharing. A new normal is in our future, where open, understood, and comparable data metrics are responsibly adopted. A significant part of the infrastructure that is needed to support this development of data metrics has been built, including ways to normalize data usage and citation, and open infrastructure to share these counts. These components are essential steps in the development of data metrics; including bibliometrics and qualitative studies, along with community buy-in, moves us closer to this bright state.

Contextualizing the counts

Infrastructure providers are not experts at analyzing the correlations and behaviors that can be found in the data citations and data usage they report. Thus, bibliometricians and data scientists who study these sorts of relationships should play an essential role in this ecosystem by developing metrics for the community. There are many questions to investigate while developing an understanding of data metrics, and the following questions are a start:

1. What is the correlation between data views and downloads? Is there a fixed relationship or are there significant differences, for example by discipline?

2. What is the pattern of data citations and data usage over time?

3. How do data citations and data usage correlate with each other? And how do they correlate with altmetrics indicators such as tweet counts and Wikipedia mentions?

4. What are the disciplinary differences in data citations and data usage?

These are all baseline bibliometrics questions that will require qualitative and quantitative assessment, and they are essential in beginning to understand what usage and citation for data mean.

Once a better understanding of these data metrics basics is achieved, specific and highly relevant topics can be investigated by using data usage and data citations to demonstrate the extent to which data sharing takes place. Another important set of questions that can be addressed involves assessing the return on investment into research supporting systems, and the effectiveness of initiatives and policies. Some example questions include:

1. Do repository certifications (such as CoreTrustSeal) [1] lead to increased reuse of the datasets they host?

2. Can it be demonstrated that policies such as the Enabling FAIR Data Commitment Statement in the Earth, Space, and Environmental Sciences [2] have a positive impact on data sharing with increased reuse and recognition?

3. Do highly cited papers have underlying data that are more frequently downloaded and cited?

4. Do training courses and workshops for data science lead to an increase in data sharing and data usage?

While bibliometrics studies do not inherently assign meaning to counts or statistics, these analyses can spark community uptake and discussion, leading to a better understanding of data metrics. The goal of involving bibliometricians is to not only address these and other important research questions but to also provide the foundation for ongoing research on research data. As adoption of open frameworks for normalized data usage and citation increases, and natural shifts in research culture as time goes on, there will be further room for analysis and studies.

Notes

1. CoreTrustSeal. (2019). Retrieved November 1, 2019, from CoreTrustSeal website: https://www.coretrustseal.org/
2. Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., … Wyborn, L. (2019). Make scientific data FAIR. *Nature*, *570*(7759), 27–29. https://doi.org/10.1038/d41586-019-01720-7

Bringing in the qualitative

Data usage and data citation cannot assess the usability, scientific accuracy, or usefulness of a dataset. Qualitative assessment needs to be included in the development of data metrics to give a broader understanding of the impact and reach of data.

One key element of this qualitative assessment is understanding the drivers of usage. [1] For example, was a dataset frequently downloaded because it was tweeted about, versus it being used for scientific research, or in teaching? To address these kinds of questions, correlations with other events, e.g. citations, or patterns over time can be considered. This can be greatly facilitated using machine learning tools and by including where users are located [2]

Another role for human assessment is evaluating the quality and usability of data through data curation, data peer review, and post-publication assessment. The value of proper and FAIR data curation is increasingly highlighted, and research supporters have shifted focus to ensuring that more data are curated before publication. Curation aids in the usability of a dataset but is not a scientific evaluation of the data itself.

Peer review is a useful component in this process. Data can be cited, tweeted, and downloaded for various reasons, but if the data are inaccurate or incomplete, this needs to be highlighted. Peer review of data is not yet standardized and is still very much in its early days. As such, calling out the need for peer review to become common practice will aid in developing more transparent science and associated metrics.

Engaging researchers, editors, curators, journals, and repositories to increase data curation and data review is a step forward that various communities can take together. This includes engaging the qualitative research communities on these more bibliometric-focused questions to ensure that metrics development does not rely strictly on usage and citation counts as indicators of researcher behavior and data impact.

Notes

1. Gordon, G., Lin, J., Cave, R., & Dandrea, R. (2015). The Question of Data Integrity in Article-Level Metrics. *PLOS Biology*, *13*(8), e1002161. https://doi.org/10.1371/journal.pbio.1002161
2. The COUNTER Code of Practice for Research Data supports optional reporting of usage by country or state, though it is limited by privacy laws in how granular this spatial information can be.

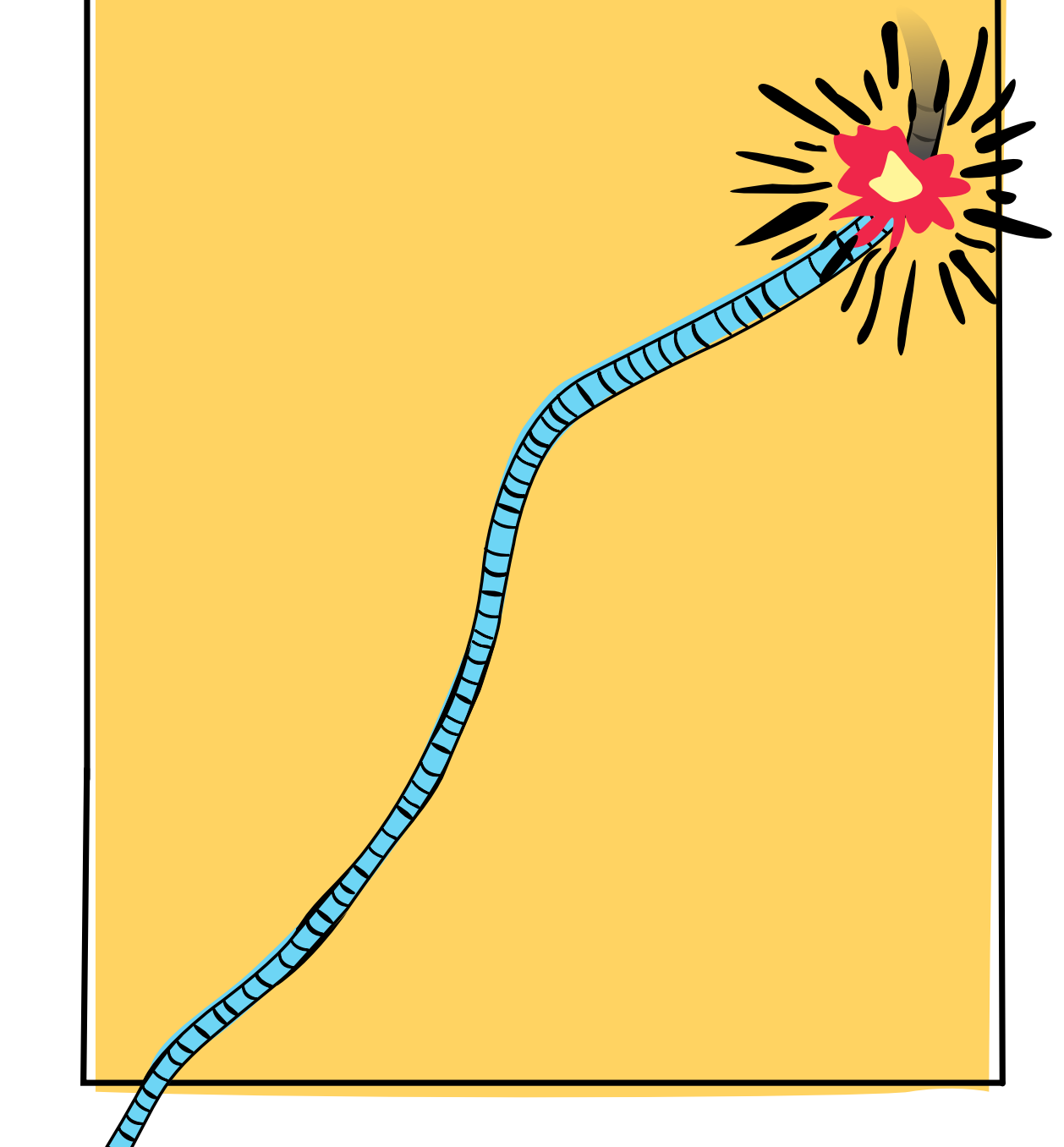
Growing a responsible community

As discussed, to achieve the vision articulated for open data metrics, it is essential that the community weigh in, adopt, and own the open framework and principles outlined in this book. With the understanding of what these numbers and correlations mean, providing qualitative oversight, and developing open data metrics, the community can then invest in this space by building tools and services that provide value.

Being that these datasets, data metrics, and infrastructure are all open, we cannot be exclusionary. In our bright future state, data metrics are a new normal. This means, for example, that data metrics are as common a topic in scholarly communications and research as journal articles. It also means that all supporters, including commercial entities, should be responsibly contributing to, and building on, these metrics without creating new systems that compete with adopted or community-owned systems.

Understanding that data metrics can alter behavior in unintended ways, the community needs to ensure that use of data metrics is not exploitative or misaligned with scientific motivations. By regularly emphasizing community input and perceived values, promoting these metrics to researchers in beneficial and realistic ways, and assessing the changing needs in the research space, open data metrics should be adaptable and truly owned by the community.

6 Lighting the fire



Research data are, and should continue to be, highly valued. Placing value on research data involves building the infrastructure to publicly archive research data, but also building community agreements on the benefits of sharing and reuse of data. Part of this includes and relies on the development of research data metrics.

The journey to data metrics starts with building on the community agreements around the *value* of research data sharing. Combining *standards* and *open infrastructure* to *normalize* approaches to counting data usage and citation is a baseline step.

With *trusted* data usage and citation information in a *centralized*, open hub, bibliometricians, data scientists, and others can begin to study the correlations, relationships and disciplinary differences within data reuse and citation. Qualitative studies understanding data publication and data usage behaviors can build on these quantitative analyses and will give us better insights into the impact and reach of research data.

Most importantly, the development of data metrics requires *community* input, iteration, and buy-in. Now that the flame has been ignited, let's set the world on fire with open data metrics.

About the authors

Daniella Lowenberg (https://orcid.org/0000-0003-2255-1869) is the Dryad Product Manager and Make Data Count lead, based at the University of California Curation Center (UC3) at the California Digital Library (CDL). Working on cross-organizational initiatives and chairing community groups such as the RDA Data Usage Metrics working group, she focuses on open source solutions for research data publishing and metrics. Prior to this, Daniella worked at PLOS ONE where she led the implementation of the PLOS Data Policy. Before joining the access and publishing side of science, she researched and published on antibiotic resistance and pharmacogenomics.

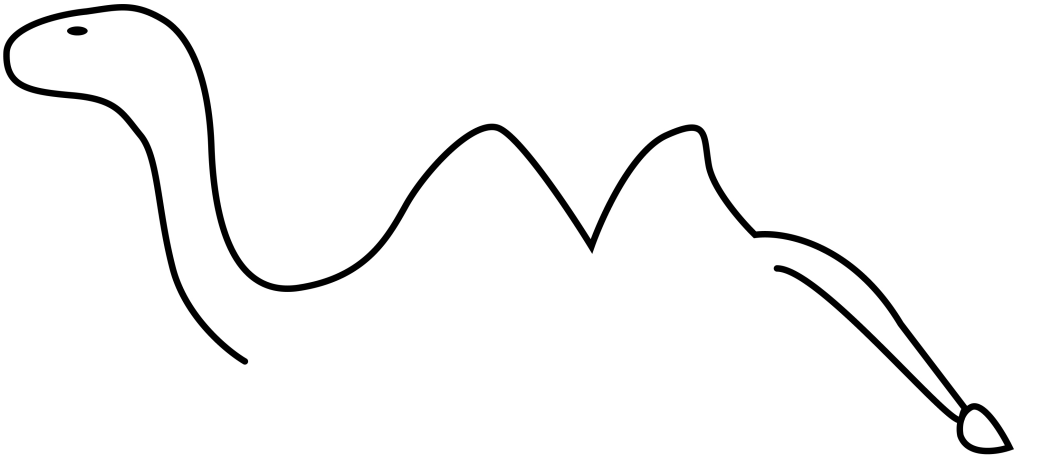
John Chodacki (http://orcid.org/0000-0002-7378-2408) is the Director of the University of California Curation Center (UC3) at California Digital Library (CDL). As UC3 Director, John works across the UC campuses and the broader community to ensure that CDL's digital curation services meet the emerging needs of the scholarly community — including digital preservation, persistent identifiers, data management, and data publishing. He currently serves on the board and/or steering committees of DataCite, FORCE11, ROR (Research Organization Registry), COUNTER, Collaborative Knowledge (Coko) Foundation, and Metadata 2020.

Martin Fenner (https://orcid.org/0000-0003-1419-2405) is the DataCite Technical Director since 2015. He is or has been involved in a number of data sharing initiatives and projects, including the RDA Scholarly Link Exchange WG, Force11 Data Citation Implementation Pilot, and

Make Data Count. From 2012 to 2015 he was the technical lead for the PLOS Article-Level Metrics project. Martin has a medical degree from the Free University of Berlin and is a Board-certified medical oncologist. He has been blogging about scholarly infrastructure since 2008. He lives in Münster, Germany.

Jennifer Kemp (http://orcid.org/0000-0003-4086-3196) is Head of Business Development at Crossref. Prior to Crossref, she worked at Springer Nature in a variety of roles and as a Publication Manager at HighWire Press. Jennifer's career experiences have exposed her to the idiosyncrasies of publishing in the breadth of science and humanities disciplines and her perspective on scholarly communications remains influenced by her years as a librarian, at IBM Research, where she was focused mainly on electronic resources acquisitions and management.

Matthew B. Jones (https://orcid.org/0000-0003-0077-4738) is Director of Informatics at the National Center for Ecological Analysis and Synthesis at the University of California, Santa Barbara. He is the Director of the DataONE program, a global network of interoperable data repositories, and of the NSF Arctic Data Center. Matt's career has focused on improving data science infrastructure to support cross-disciplinary and synthetic science, principally through the development of open source software for data repositories, metadata systems, and reproducible analysis and modeling.



Original drawing during the planning phase of the book attempting to equate open data metrics to a two-humped camel. Enjoy

Colophon

The authors would like to acknowledge the Alfred P. Sloan Foundation for funding the writing of this book. The first draft of this book was written in a five-day Book Sprint with the Book Sprints methodology (www.booksprints.net).

Book Sprints facilitation: Faith Bosworth

Copy editing: Raewyn Whyte and Christine Davis

HTML book design: Manuel Vazquez

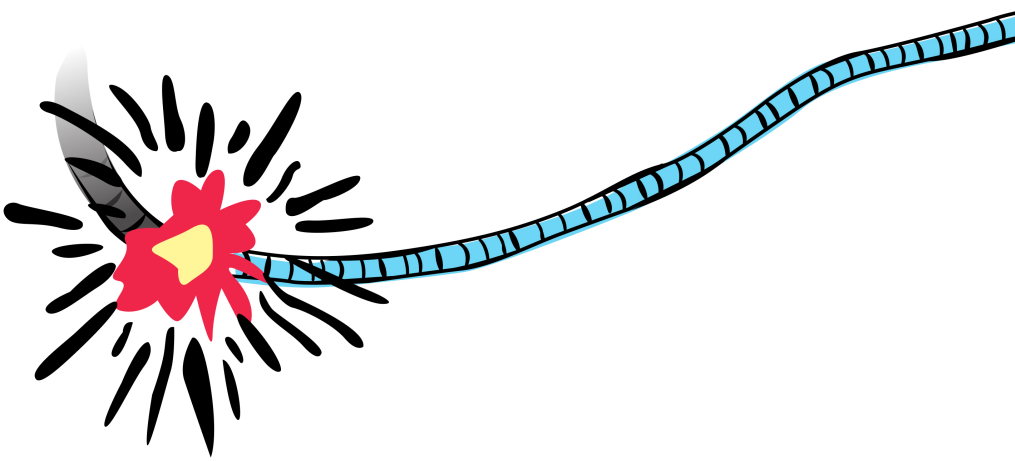
Illustrations and cover design: Henrik Van Leeuwen and Lennart Wolfert

Fonts: Barlow (designed by Jeremy Tribby); Roboto (designed by Christian Robertson); Bariol (designed by Atipo Foundry)

Cite as: Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., Jones, M. (2019). *Open Data Metrics: Lighting the Fire (Version 1).* Zenodo. https://doi.org/10.5281/zenodo.3525349

Research data is at the center of science, and to date it has been difficult to understand its impact. To assess the reach of open data, and to advance data-driven discovery, the research and research supporting communities need open, trusted data metrics.

In *Open Data Metrics: Lighting the Fire*, the authors propose a path forward for the development of data metrics. They acknowledge historic players and milestones in the process and demonstrate the need for standardized, transparent, community-led approaches to establish open data metrics as the new normal.



Lowenberg, D., Chodacki, J., Fenner, M., Kemp, J., Jones, M. (2019). *Open Data Metrics: Lighting the Fire (Version 1)*. Zenodo. http://doi.org/10.5281/zenodo.3525349