Scholarly Assessment Reports

Konkiel, S. (2020). Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators. *Scholarly Assessment Reports*, 2(1): 13. DOI: https://doi.org/10.29024/sar.13

COMMISSIONED REPORT

Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators

Stacy Konkiel
Altmetric, US
stacy@altmetric.com

Research data in all its diversity—instrument readouts, observations, images, texts, video and audio files, and so on—is the basis for most advancement in the sciences. Yet the assessment of most research programmes happens at the publication level, and data has yet to be treated like a first-class research object.

How can and should the research community use indicators to understand the quality and many potential impacts of research data? In this article, we discuss the research into research data metrics, these metrics' strengths and limitations with regard to formal evaluation practices, and the possible meanings of such indicators. We acknowledge the dearth of guidance for using altmetrics and other indicators when assessing the impact and quality of research data, and suggest heuristics for policymakers and evaluators interested in doing so, in the absence of formal governmental or disciplinary policies.

Policy highlights

- Research data is an important building block of scientific production, but efforts to develop a framework for assessing data's impacts have had limited success to date.
- Indicators like citations, altmetrics, usage statistics, and reuse metrics highlight the influence of research data upon other researchers and the public, to varying degrees.
- In the absence of a shared definition of "quality", varying metrics may be used to measure a dataset's accuracy, currency, completeness, and consistency.
- Policymakers interested in setting standards for assessing research data using indicators should take into account indicator availability and disciplinary variations in the data when creating guidelines for explaining and interpreting research data's impact.
- Quality metrics are context dependent: they may vary based upon discipline, data structure, and repository. For this reason, there is no agreed upon set of indicators that can be used to measure quality.
- Citations are well-suited to showcase research impact and are the most widely understood indicator. However, efforts to standardize and promote data citation practices have seen limited success, leading to varying rates of citation data availability across disciplines.
- Altmetrics can help illustrate public interest in research, but availability of altmetrics for research data is very limited.
- Usage statistics are typically understood to showcase interest in research data, but infrastructure to standardize these measures have only recently been introduced, and not all repositories report their usage metrics to centralized data brokers like DataCite.
- Reuse metrics vary widely in terms of what kinds of reuse they measure (e.g. educational, scholarly, etc). This category of indicator has the fewest heuristics for collection and use associated with it; think about explaining and interpreting reuse with qualitative data, wherever possible.
- All research data impact indicators should be interpreted in line with the Leiden Manifesto's principles, including accounting for disciplinary variation and data availability.
- Assessing research data impact and quality using numeric indicators is not yet widely practiced, though there is generally support for the practice amongst researchers.

Keywords: Altmetrics; citations; reuse; usage statistics; quality indicators; research data

Introduction

Research data is the foundation upon which innovation and discovery are built. Any researcher who uses the scientific method—whether scientist or humanist—typically generates data. Much of this data is analyzed by computational means and archived and shared with other researchers to ensure the reproducibility of their work and to allow others to repurpose the data in other research contexts.

Given the importance of data, there have increasingly been calls for research data to be recognized as a first-class research object (Force11, 2015; National Information Standards Organization, 2016; Starr et al., 2015), and treated as such in research evaluation scenarios. Understanding research impact through the lens of publication and patent analysis is nowadays commonplace; less is understood about how to evaluate the quality and impact of research data.

In this article, we discuss how research indicators can be used in evaluation scenarios to understand research data's impact. We begin with an overview of the research evaluation landscape with respect to data: why policymakers and evaluators should care about research data's impact; the challenges of measuring research data's quality and impact; and indicators available for measuring data's impact. We then discuss how to use data-related indicators in evaluation scenarios: specifically, how data metrics are being used in current evaluation scenarios, and guidelines for using these metrics responsibly.

The goals of this article are to provide the community with a summary of evidence-backed approaches to using indicators in assessment of research data and suggest heuristics that can guide evaluators in using these metrics in their own work.

What are research data?

Merriam-Webster defines data as "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation".[1] Research data is data used in the service of empirical analysis in the sciences and humanities. The University of Queensland policy on the management of research data[2] explains further:

> "Research data means data in the form of facts, observations, images, computer program results, recordings, measurements or experiences on which an argument, theory, test or hypothesis, or another research output is based. Data may be numerical, descriptive, visual or tactile. It may be raw, cleaned or processed, and may be held in any format or media."

Crucially, research data is analyzed as the basis for new discovery and is not the results of an analysis (e.g. figures or tables). Data can be observational (unique data collected in real time), experimental (data collected under controlled conditions, often using lab equipment), simulation (data generated from test models), compiled (data combined from existing analyses or data sources like texts), or reference (curated, peer review data compiled by organizations) (Claibourn, n.d.).

Research data exists in many different digital formats. These can include photographs, geospatial data, videos, text, audio files, databases, instrumentation outputs.

Specific definitions for what constitute "data" and popular formats tend to vary by field (**Table 1**). There are a wide array of dataset publishers and a number of services exist to provide dataset metrics. **Table 2** provides a brief overview of the major players in dataset infrastructure; see the Registry of Research Data Repositories[3] for a comprehensive list of dataset repositories, specifically.

At a high level, datasets are defined by the COUNTER Code of Practice for Research Data 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" (Project COUNTER, n.d.-b). A dataset may be comprised of components (smaller parts of a dataset that can be downloaded individually), and may have different versions (fixed representations of a dataset at a particular point in time, or including specific data features like columns) (Lowenberg et al., 2019).

Table 1: Examples of research data, by discipline.

Discipline	Data types	Data formats
Biology	DNA sequences, microscopy images, morphological data, images of specimens	Images (.jpg, .tiff), FASTQ (.fq), SAM (.sam)
Sociology	Survey responses, digitized videos, labor statistics, economic indicators	Spreadsheets (.csv, .xlsx), videos (.mov, .mp4, .avi), text files (.txt, .xml, .docx), databases (.sql, .csv)
History	Speech transcripts, digitized photographs and videos, journals, newspaper clippings, diaries	Text files (.txt, .xml, .docx), images (.png, .jpg, .tiff), videos (.mov, .mp4, .avi)

[1] https://www.merriam-webster.com/dictionary/data.

[2] http://www.mopp.qut.edu.au/D/D_02_08.jsp.

[3] https://www.re3data.org/.

Table 2: Overview of major dataset repositories and services.

Service	Repository	Description
Repository (Generalist)	Zenodo	A nonprofit, open access repository serving all subject areas. Administered by CERN.
Repository (Subject)	Dryad	A nonprofit, open access data repository serving all subject areas, with special emphasis on evolutionary, genetic, and ecology biology.
Repository (Generalist)	Figshare	A commercial, open access repository serving all subject areas. A subsidiary of Digital Science.
Repository (Subject)	Inter-university Consortium for Political and Social Research (ICPSR)	A nonprofit data archive and repository serving the social sciences. Administered by the Institute for Social Research at the University of Michigan.
Metadata; Metrics Provider (Citations, Usage)	DataCite	A nonprofit dataset registry, DOI service, metadata index, and metrics provider.
Metrics Provider (Citations)	Data Citation Index	A commercial service that indexes datasets, data publications, and citations to data in the research literature. A subsidiary of Clarivate Analytics.
Metrics Provider (Altmetrics)	Altmetric	A commercial service that tracks research shared online, including datasets, and makes both quantitative and qualitative engagement and discussion data ("altmetrics") available. A subsidiary of Digital Science.
Metrics Provider (Altmetrics)	PlumX Metrics	A commercial service that tracks research shared online, including datasets, and makes both quantitative and qualitative engagement and discussion data ("altmetrics") available. A subsidiary of Elsevier.
Repository Registry	Registry of Research Data Repositories	A service operated by the nonprofit DataCite, offering information on more than 200 research data repositories worldwide.

How one defines research data has a bearing on how data is stored and accessed. For example, the contents of historical speech transcripts would not change over time and might be used regularly in teaching, whereas a collection of bird species images could be updated regularly to reflect new discoveries and shared widely by birding enthusiasts on social media. Differences in data storage and access can affect data's use and the related metrics that can evaluate the data's impacts (Lowenberg et al., 2019).

Evaluation of research data

There have been a number of changes in recent years that shape why and how evaluators may assess research data:

- The volume of research data available worldwide has grown with advances in the ease of creating "born digital" research and using high-capacity computing resources (Anderson, 2008; Costas et al., 2013).
- Developments in "citizen science" have raised concerns about dataset quality for data collected, maintained, and described by members of the public (Exel et al., 2010; Leibovici et al., 2017).
- The research community has increasingly shown support for Open Access, and along with it "Open Research" practices like data sharing (Cabello Valdes et al., 2017; Laakso & Björk, 2012; Piwowar et al., 2018; Piwowar & Vision, 2013).
- The development of the "data publication", an article-like format for sharing descriptions of datasets; these may make datasets into more "citable" formats (Ingwersen & Chavan, 2011; Leitner et al., 2016).
- The recent rise of altmetrics, a class of scientometric indicators that help evaluators understand the attention that research has received online (Moed, 2016), and new developments in citation-based metrics, have made it possible to understand the impacts of data among new audiences like the public and policymakers (Konkiel, 2013; Peters et al., 2016; Piwowar, 2013).

It is in this context that the evaluation community has been considering how to use indicators to understand the quality and impact of research data. There is also an interest in using indicators to incentivize data sharing—the idea being that the increased use of data-related indicators in evaluation practices would encourage researchers to begin sharing their data more often (Bierer et al., 2017; Borgman, 2012; Konkiel, 2013).

Caveats to using research data indicators

Before we discuss the many indicators that can be used to understand the impacts of research data, it is important to consider the challenges that exist to developing useful and sound metrics, given various socio-technical limitations. Note that throughout this article, we differentiate between *indicators* (numerical measures that are calculated and interpreted to quantify particular outcomes like productivity) and *metrics* (numerical measures that simply report web-based events like downloads).

Data is often dynamic

Perhaps the biggest challenge to understanding the impact of research data is that data can change. As new observations are gathered, data may be updated and "versioned", making a dataset at one point in time potentially very different from the same dataset a year later. Moreover, one version of a dataset may be substantially different in content from others, in terms of the volume of data collected or the features included in the dataset (Lowenberg et al., 2019). In this way, data is different than publications like journal articles or monographs, which are generally understood to be fixed, "point in time" objects describing research (notwithstanding preprints and other versions of a work that goes on to be published in a peer reviewed journal or by a scholarly press).

Where data is in flux, it can be difficult to compare citation rates or altmetrics for research data generated in the same year, or even comparing a dataset's citations in one year to the next. Thus, time-bound normalized indicators (e.g. comparing all datasets published in the same subject area and same year), which are typically used to account for variance, should not necessarily be used to make comparisons against changing data.

Metrics providers themselves also introduce challenges, by way of their product design choices and how they represent different versions of a single research output like a dataset. For example, Figshare, Altmetric, and PlumX Metrics all report metrics for various versions of a dataset in a single item record, with no differentiation made between attention received by different versions of the dataset. These design choices make it difficult to tease out the reuse and attention for a dataset at a particular point in time.

Measuring reuse is difficult

Data is often shared with the express purpose of allowing others to reuse and repurpose the data to fuel new discoveries. This purposeful atomization of research objects makes it difficult to track the cumulative impact of a single dataset (Lowenberg et al., 2019; Missier, 2016), especially in cases where repositories have not implemented standards for ingesting and sharing usage metrics and other indicators.

These challenges are due to a number of technical barriers. In our current research environment, data citation standards vary from field to field (**Table 3**), and research data can be shared on a variety of platforms (from trusted repositories like Dryad to researcher websites and Google Drive) (Altman et al., 2015). Moreover, many current metrics platforms are not designed to accurately track and collate attention in a manner that maps the impacts of derivative data back to the original dataset (in other words, to measure reuse) (Lowenberg et al., 2019). Provenance metadata, which offers a mechanism to track the reuse of data across studies, has been implemented in metadata provider DataCite and in the SCHOLIX standard (Burton et al., 2017) but has not yet been leveraged by any major dataset metric provider to provide reuse indicators.

Variance among metrics provider data sources

Though many data providers share metrics and indicators that can be used to track the impacts of data, these providers often rely upon different data sources that are not directly comparable or collect metrics from the same set of providers in different ways.

Citation counts can vary between providers, due to differences in where the services track for citations. The Data Citation Index offers a comprehensive view on citations to research data and data publications within the peer-reviewed research literature, sourced from Web of Science, the Chinese Science Citation Database, and other Clarivate Analytics-owned citation databases, in addition to data from partners like SciELO (Rivalle & Green, 2018). Its closest competitor, DataCite's data metrics API, includes similar citations, which are sourced from publications indexed by Crossref and reported by the publisher, or sourced from related datasets indexed in and reported by DataCite's partner repositories (*Datacite Citation Display*, n.d.).

Similarly, altmetrics services treat research data differently from one another, and often index different attention sources. Altmetric treats datasets as they do any other research object, indexing files that are assigned persistent

Table 3: Examples of data sharing norms, by discipline.

Discipline	Shares data?	Explanation
Astronomy	Usually	"The fact that astronomical data from large surveys are publicly available is remarkable, but by no means surprising. Astronomers collect data about the Universe, and thus, they may feel a moral obligation to share collected data openly." (Pepe et al., 2014)
Political Science	Sometimes	Concerns over sharing personally identifiable or sensitive information, particularly for qualitative data (e.g. interviews) (Buthe et al., 2015)
Medical Sciences	Sometimes	HIPAA protections of patient data (Freymann et al., 2012); anonymization makes data sharing possible (Hrynaszkiewicz et al., 2010)

identifiers and shared in one of the 17 sources they track. Plum Analytics also takes this approach, and supplements it by providing data-specific metrics from data sharing platforms. Comparative analyses have shown differences in altmetrics between these two providers (Enkhbayar et al., 2020; Meschede & Siebenlist, 2018; Peters et al., 2016, 2017; Zahedi et al., 2015).

Moreover, metrics services sometimes collect data from the same sources in different ways. For example, two providers may have different standards for what "counts" as a citation to a dataset, with one tracking any link in a document as a citation, and another only counting a citation if it appears in the references list of a peer-reviewed journal article. Similarly, altmetrics services may track a source like Facebook very differently: for the sake of "auditability," Altmetric tracks only links to research that appear in posts on public Facebook pages, while Plum Analytics counts Facebook likes, comments, and shares across the entire platform, including for private posts (Zahedi et al., 2015). These nuances are not always apparent to the end user, despite the best efforts of metrics providers.

Platform-specific metrics

Research data is shared across hundreds of repositories worldwide, each capturing and reporting metrics with varying degrees of transparency, granularity, and completeness. Reporting usage statistics like downloads to DataCite is voluntary,[4] with overall coverage rates unknown. Altmetrics providers track only a portion of available data repositories (with Figshare, Dryad, and Pangea being the best-known examples). Typically, this is due to platform design limitations that make tracking data repositories difficult, for example through the inconsistent use of webpage meta tags (Gregg et al., 2019).

Overall, the prevalence of platform-specific metrics and the dearth of reporting to and tracking by centralized metrics providers like DataCite and Plum Analytics make it difficult to accurately benchmark or create normalized metrics for dataset usage.

Available research data indicators

Despite these challenges, there are a number of research data indicators that evaluators can use to support their assessment of the quality and impact of datasets. These indicators can be categorized into five main classes: quality indicators, citation-based indicators, altmetrics, usage statistics, and reuse indicators. Here, we will describe each class of indicator, paying special attention to their strengths and limitations for evaluation scenarios.

Quality indicators

A dataset's quality can be understood as its accuracy, currency, completeness, and consistency (Fox et al., 1994), in addition to many other possible dimensions, which theorists have settled upon in the absence of a common definition for "quality" itself. For full understanding of possible data quality dimensions, see (Sidi et al., 2012; Wand & Wang, 1996).

Quality dimensions typically have associated measures, from which indicators of a dataset's quality can be derived. For example, Fox *et al* (1994) suggest that accuracy can be measured by the size of a dataset's error, e.g. the fraction of incorrect values in a dataset. **Table 4** includes examples of dataset quality dimensions and their indicators; a comprehensive list can be found in Batini et al. (2009).

Pipino et al. (2002) suggest that dataset quality indicators fall into three main categories: simple ratios (e.g. an objective measure like Fox's fraction of incorrect dataset values), min or max operations (often a subjective ranking where a

Table 4: Data quality dimensions and indicators, adapted from Batini et al., 2009.

Dimensions	Possible Measures
Accuracy	Syntactic Accuracy=Number of correct values/number of total values User Survey – Questionnaire
Currency	Currency = Time in which data are stored in the system – time in which data are updated in the real world Time of last update Currency = Request time – last update Currency = Age + (Delivery time – Input time) User Survey – Questionnaire
Completeness	Completeness = Number of not null values/total number of values Completeness = Number of tuples delivered/Expected number Completeness of Web data = (Tmax – Tcurrent)∗ (CompletenessMax – CompletenessCurrent)/2 User Survey – Questionnaire
Consistency	Consistency = Number of consistent values/number of total values Number of tuples violating constraints, number of coding differences Number of pages with style guide deviation User Survey – Questionnaire

[4] https://support.datacite.org/docs/eventdata-guide.

dataset is rated on a scale), and a weighted average (where similar dimensions' ratings are factored together to create a single indicator).

Indicators can be helpful in statistical monitoring for data quality, especially for clinical trials and other disciplines that produce highly structured data (George & Buyse, 2015; Knepper et al., 2016). These fields can use tests that measure a dataset's completeness, variability in measurement errors, and other indicators to automate quality checks (Knepper et al., 2016). Examples of automated data quality checks can be found on the data science platform, Kaggle: the "Usability" score helps users understand the extent to which a dataset is documented and how often it is updated, and column-wise summaries highlight missing (null) values.[5]

However, it is unclear to what extent automated statistical monitoring are helpful in fields that produce unstructured and semi-structured data. In such disciplines, manual data quality checks are an important safeguard against fraud. Some repositories like ICPSR perform manual data quality checks as part of the data deposit process;[6] therefore, inclusion in such a repository can be thought of as an indicator of quality, on its own.

It is important to note that the context of a dataset can shape its data quality dimensions and related indicators. For example, open geospatial data quality has distinct dimensions regarding openness, mapping capabilities, and other features that factor into concepts of a dataset's quality; these dimensions would not necessarily appear in other kinds of datasets like polling data (Xia, 2012). Stausberg *et al* (2019) describe this phenomenon as "contextual quality" (Stausberg et al., 2019).

Dataset quality measurement is an ongoing, iterative endeavor. As Pipino et al. write (2009), "assessing data quality is an ongoing effort that requires awareness of the fundamental principles underlying the development of subjective and objective data quality metrics."

Citation-based indicators

Citations are used widely to understand the scholarly impact of research publications like journal articles, monographs, and edited volumes, especially in the sciences (Bornmann & Daniel, 2008; De Rijcke et al., 2016). In recent years, researchers have suggested that data should be cited as a "first-class research output" when used as the basis for subsequent studies (Callaghan et al., 2012; CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013; Lawrence et al., 2011; Silvello, 2018). Researchers may either cite the datasets themselves, or "data publications"—summaries of dataset scopes, formats, and other such information needed to reuse data.

Citations to research data can be interpreted in a number of ways:

- A signal for **value**, as researchers base their own studies upon data created by others, or replicate studies using open data (Mooney & Newton, 2012; Piwowar & Vision, 2013; Silvello, 2018),
- An indicator of **influence upon a discipline** (Fear, 2013), and
- A measure of **reuse** by other researchers (Hahnel, 2013)

For a complete overview of the many possible meanings of data citations, consult Silvello's "Theory and practice of data citation" (2018).

Data citations function similarly to traditional citations in academic texts, providing a way to acknowledge intellectual debt to scholarly forebears. They also allow scientists to connect research articles to the data they are based upon and give credit to others who have shared their data openly (Silvello, 2018).

Dataset citations can be referenced in texts similarly to how monographs and journal articles are referenced, typically by noting the authors and year of publication. The specific formatting of a dataset reference will differ, depending upon the preferred writing style of the publisher (**Table 5**).

The Data Citation Index (Clarivate Analytics) and DataCite are two of the best-known providers of data citation indicators. The Data Citation Index (DCI) has indexed more than 2.6 million records, tracking over 400,000 citations to

Table 5: Examples of a dataset citation, by writing style.

Writing style	Reference
APA (6th Edition)	Powers, J. et al. (2020). *A catastrophic tropical drought kills hydraulically vulnerable tree species* [data file and publication]. Dryad [distributor]. doi: 10.5061/dryad.2rbnzs7jp
Chicago (16th Edition)	Powers, Jennifer et al. 2020. *A catastrophic tropical drought kills hydraulically vulnerable tree species.* Distributed by Dryad. doi: 10.5061/dryad.2rbnzs7jp
DataCite	Powers, Jennifer et al. (2020), A catastrophic tropical drought kills hydraulically vulnerable tree species, v4, Dryad, Dataset, https://doi.org/10.5061/dryad.2rbnzs7jp

[5] https://www.kaggle.com/.

[6] A list of repositories that offer data quality checks can be found on re3data.org: https://www.re3data.org/search?query=&qualityManagement%5B%5D=yes.

datasets, data papers, and data repositories (Robinson-Garcia et al., 2015) in a searchable database. The DCI covers a range of subject areas, with particular strength in the sciences (Robinson-Garcia et al., 2015). Citations to datasets are sourced from the the DCI, Web of Science, the Chinese Science Citation Database, and other Clarivate Analytics-owned citation databases, as well as from partners like SciELO (Rivalle & Green, 2018).

DataCite is a content registration service, and its Event Data API provides citation indicators for research outputs that link to DataCite records (*DataCite Event Data*, n.d.). DataCite has handled over 6 million DOI registrations to date, of which 42% were specifically for datasets (Robinson-Garcia et al., 2017). A sizable portion of DataCite-registered content comes from only a handful of repositories, and due to inconsistency in metadata, it is unclear whether there is disciplinary bias within the content registered via DataCite (Robinson-Garcia et al., 2017). DataCite citation indicators are currently only available via API, requiring users to be technically proficient. As of February 2020, around 450,000 citations were reported via the DataCite API.

Citation-based indicators are well-suited for evaluation scenarios for a few reasons. The practice of citing—whereby an author references publications or other resources that have influenced their own study—is a concept that is understood and valued by many within academia, and thus is relatively easy for evaluators to interpret.

Given their legibility, citations can also help evaluators to better understand the value of "Open" research practices. Citations can help evaluators and decision-makers understand the value of sharing data for others to reuse (Drachen et al., 2016). Data citations can also draw attention to data curation as an act of scholarship (Belter, 2014; Silvello, 2018).

However, data citations should be interpreted carefully, due to the lack of coverage across disciplines (Mongeon et al., 2017). A deeply engrained authorship and citation culture affect the rates at which research data is cited: some journals and data repositories lack standardized guidelines for citing data (Mooney & Newton, 2012), informal data citation is more common in certain fields (Park et al., 2018; Zhao et al., 2018), and authors may not be aware of the need to cite the data they base their studies upon (Mooney & Newton, 2012). There are also known differences between the types of datasets and their citation rates (Peters et al., 2016).

Technical barriers also exist to the widespread use of data citation in evaluation. The proliferation of scholarly identifiers is an ongoing challenge that can make it difficult to reference datasets in a consistent and stable manner ("On the Road to Robust Data Citation," 2018), which in turn makes it harder to accurately track citations. Many evaluators also face a dearth of citation data upon which to base their evaluations, or verify submitted citation figures—the Data Citation Index and DataCite are not yet widely adopted in evaluation contexts, and in the humanities, the practice of data citation is rare (Peters et al., 2016).

With these strengths and limitations in mind, it is possible that evaluators and policymakers who are highly attuned to disciplinary data citation practices can potentially use data citations to understand the impacts of the research they are evaluating.

Altmetrics

Altmetrics are data that highlight how research has been shared, discussed, or otherwise engaged with online. They are collected when a research output is linked to or mentioned in a source that altmetrics aggregators track. In recent years, altmetrics have increasingly been suggested as a means of understanding the broader impacts of research data. However, these highly heterogeneous data can mean many different things, and should be interpreted carefully, usually in tandem with other indicators (Konkiel, 2016).

For the purposes of this article, we define altmetrics as links to research from online platforms and documents that add commentary and value when research is shared, e.g. public policy, social media, peer reviews, patents, and software sharing platforms. Sugimoto *et al* (2017) have further defined scholarly social media use as social networking, social bookmarking and reference management, social data sharing, video, blogging, microblogging, wikis, and social recommending, rating, and reviewing.

Altmetrics are distinct from usage statistics and other webometrics like referral links, and from citation-based indicators (Glänzel & Gorraiz, 2015). Certain altmetrics like GitHub-based metrics can also be considered reuse metrics.

Altmetrics are typically interpreted as a proxy for:

- **Social impact** (Bornmann, 2014; Erdt et al., 2016)
- **Attention or "buzz"** (Sugimoto, 2015; Thelwall et al., 2013)
- **Reach or readership** (Dinsmore et al., 2014; Konkiel & Scherer, 2013; Mounce, 2013)
- In rare cases, **quality** (Bornmann & Haunschild, 2018; Nuzzolese et al., 2019; Sud & Thelwall, 2014)

Altmetrics are often praised for their ability to reflect engagement with research on a much faster timescale than citations can—it can take hours for the first mentions of a dataset to be tracked by an altmetrics aggregator, while it typically takes months for citations to appear in the peer-reviewed literature. For example, altmetrics for coronavirus-related research increased rapidly starting in mid-February 2020, shortly after the virus became a global public health concern (**Figure 1**).

Altmetrics can also highlight engagement with research from a broad set of stakeholders, including members of the public, science communicators, policymakers, educators, and researchers.

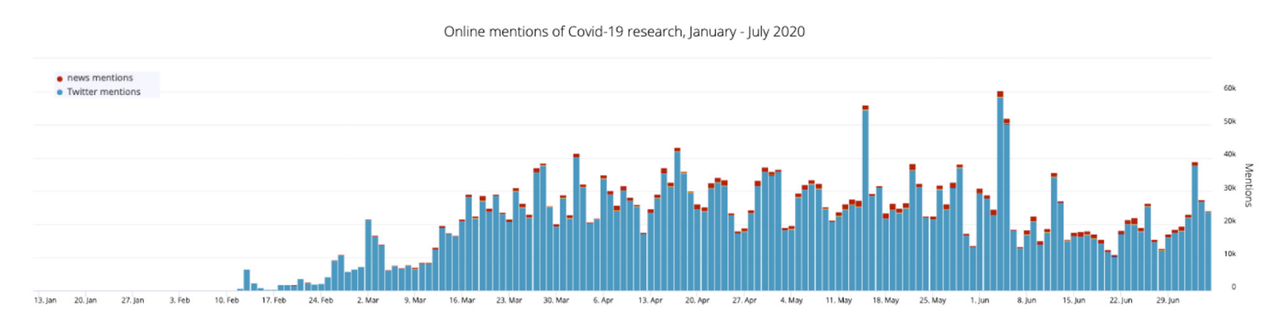

Figure 1: Online attention for research outputs with "covid-19" in the title or abstract (n = 27,824), January 2020–July 2020. A majority of online attention originates from Twitter. (data source: Altmetric Explorer).

However, much that is known about altmetrics and their meanings, stakeholders, and time-bound usage comes from research that examines engagement with journal articles. Far less is known about the communities that engage with research data on the web, beyond overall rates of engagement for cited data (Peters et al., 2016). More studies are needed.

Altmetric and PlumX Metrics are two of the best-known altmetrics services that track altmetrics for research data. These services incorporate different data sources[7,8] and in some cases use different approaches to tracking content. Across both services, less than 10% of research data has associated altmetrics (Peters et al., 2016).

Altmetric indexes more than 30,000 datasets sourced from repositories and data journals like Figshare, Dryad, and Gigascience. Altmetric treats research data similarly to all other research outputs: it tracks links to datasets that are mentioned in any of the 17 sources that the company tracks, and does not track data-specific indicators. Though Altmetric indexes research of all disciplines, it has higher rates of coverage overall in the sciences, and primarily indexes multidisciplinary and science-specific data repositories. Of Altmetric-indexed content labeled as a "data set", around 92.9% of attention occurs on Twitter, followed by the now-defunct Google+ platform (2.1%) and blogs (2%). Around 41% of dataset tweets (n = 65,909) are original (i.e. not retweets).

PlumX Metrics has collected metrics for over 450,000 datasets sourced from repositories like Dryad and Figshare, as well as data sets indexed in various institutional repositories such as Digital Commons and DSpace (S. Faulkner, personal communication, February 19, 2020). Metrics are derived from over 50 sources and organized into 5 categories – Citations, Usage, Captures, Mentions and Social Media.[9] Of all PlumX Metrics-indexed research outputs labeled as a "data set" that have associated indicators, 15.5% have Captures (primarily from Mendeley), 64.5% have Mentions (primarily from Wikipedia), 3.2% have Usage (primarily Figshare, Dryad and Digital Commons), 12.7% have Citations (primarily from CrossRef) and 4.2% have Social Media metrics (Twitter and Facebook) (S. Faulkner, personal communication, February 19, 2020).

Independent studies have found that PlumX Metrics has indexed 16% of datasets shared on Zenodo (Peters et al., 2017), and between 4% and 9% of datasets indexed in the Data Citation Index (Peters et al., 2016). For Zenodo-hosted datasets indexed in PlumX Metrics, around 6.2% of attention comes in the form of "captures" (e.g. bookmarks, "forks", and favorites), 7% from social media mentions (e.g. shares, likes, and tweets), and 1.1% from usage statistics (e.g. Figshare and Dryad views and downloads) (Peters et al., 2017). Less than 1% of Zenodo-hosted data sets received citations (sourced from citation indices such as Scopus and Crossref) or mentions (e.g. blog posts, comments, or Wikipedia articles) (Peters et al., 2017).

The company recently shared that it will soon index over 13 million datasets from more than 1,700 repositories that are shared in Mendeley Data Search (S. Faulkner, personal communication, February 19, 2020), and that usage indicators from Mendeley Data Repository will reportedly be added to PlumX Metrics as a new metrics source.

Though altmetrics are promising, there are limitations to their use in research data evaluation scenarios. The largest challenge is that altmetrics are not yet widely used in research assessment, and as such literacy surrounding their interpretation and responsible use is low (Desanto & Nichols, 2017; Kratz & Strasser, 2015). While educational resources like the Metrics Toolkit are gaining in popularity and can help inform evaluators, overall levels of knowledge remain low.

The relative ease with which online attention for research can be fabricated or "gamed" is another concern often expressed by evaluators and researchers alike (Adie, 2013; Konkiel, 2016; Roemer & Borchardt, 2015). While altmetrics services rarely see cases of intentional gaming, automated "bots" can and do artificially inflate altmetrics (Haustein et al., 2016), including for research data (Lowenberg et al., 2019), obscuring organic, meaningful engagement.

[7] For more information on Altmetric's data sources, see https://www.altmetric.com/about-our-data/our-sources/.

[8] For more information on PlumX Metrics' data sources, see https://plumanalytics.com/learn/about-metrics/.

[9] https://plumanalytics.com/learn/about-metrics/.

For example, from 2010 to 2016 the Twitter account @datadryadnew automatically tweeted all new datasets added to the Dryad repository, accounting for over 7,200 mentions of research over six years. Adie (2013) characterizes this kind of gaming as "incidental" and suggests that while it is not particularly valuable in an impact evaluation context, it can be an indicator of overall reach for research.

Altmetrics aggregators' data coverage is affected by their organizational contexts. PlumX Metrics, owned by Elsevier (and by EBSCO before that), tracks mentions from both open and proprietary data sources like SSRN, be press, and EBSCO, and has exclusive access to the latter. Figshare, a sister company to Altmetric under the Digital Science banner, and Figshare-hosted repository ChemRxiv account for more than twice the number of datasets in Altmetric than independent repositories like Dryad.

Altmetrics aggregators cannot yet offer useful disciplinary benchmarking for research data performance, due to the limited number of repositories they index. While the Data Citation Index includes hundreds of data repositories by design, Altmetric and PlumX index a much smaller number of repositories. Moreover, while research data is often still shared as supplementary files accompanying journal articles, no major altmetrics service currently records altmetrics for these files (with the exception of those files hosted by Figshare, in partnership with publishers).

Evaluators interested in understanding the social impacts of research data would be well-served by altmetrics. However, any assessment programme that incorporates these indicators should plan to develop an adaptive and iterative evaluation strategy that can address the above-mentioned caveats, because the use of research data altmetrics is still relatively new.

Usage statistics

Data usage is "counted as the accesses of a dataset or its associated metadata" (Lowenberg et al., 2019). Usage statistics for research data are usually comprised of dataset and file downloads and abstract or documentation downloads and views (Project COUNTER, 2018).

Typically, usage statistics are interpreted as:

- A proxy for **interest** in research data (Ingwersen & Chavan, 2011; Kratz & Strasser, 2015; Lowenberg et al., 2019)
- **Reuse** beyond cases where the data is incorporated into a study, e.g. downloading training data for a machine learning model (Fear, 2013)
- In some cases, an indicator for data's **use** in scientific studies (Cousijn et al., 2019; Lowenberg et al., 2019)

Usage statistics are often reported through a standard called COUNTER, which ensures consistency in how these data are tracked and reported (COUNTER, 2014). DataCite offers repositories a centralized way to share COUNTER-compliant usage data (Project COUNTER, 2018). For more information on COUNTER-compliant research data usage statistics, interested readers can consult the "Code of Practice for Research Data Usage Metrics" (Project COUNTER, 2018).

Usage statistics are often reported within data repositories, at the item record level. The extent to which these usage statistics are COUNTER-compliant varies, and is dependent upon the data repository.

DataCite is the largest centralized source for usage statistics for data repositories, indexing 6.7 million datasets from over 1900 repositories worldwide. However, not all repositories register their data with DataCite, and those that do may not share usage reports with DataCite. DataCite usage statistics are freely available via the DataCite Search and DataCite's Event Data API. As of February 2020, DataCite reports over 17.6 million views to dataset documentation, and 2.4 million downloads of the datasets themselves.[10]

Usage statistics are generally thought to help end users understand the overall interest or readership in research (Kurtz & Bollen, 2010). As Haustein (2014) explains, "Download and click rates *estimate* readership; they do not *measure* it." COUNTER-compliant usage indicators are widely accepted to be reliable, as these standards account for bots and other behaviors that can artificially inflate metrics (Project COUNTER, n.d.-a). Usage statistics are reportedly second only to citations, in terms of importance to researchers for understanding data impact (Project COUNTER, 2018).

However, usage statistics are limited in their ability to help evaluators understand who is accessing content, or how the downloaded content is being used (Kurtz & Bollen, 2010). Usage statistics that are not COUNTER-compliant are not guaranteed to be accurate, and may be unduly influenced by bots or other inorganic traffic (Lagopoulos et al., 2017; Project COUNTER, n.d.-a).

Perhaps the biggest challenge for evaluators interested in using usage statistics to understand research impact is that benchmarking is not yet widely used to make comparisons between datasets or across disciplines.

Overall, usage statistics can provide a good indicator of absolute reach and interest in a dataset, but they are limited with regard to the ability to make comparisons between datasets or understand how research data is being used.

[10] Statistics retrieved from the DataCite Events Data API https://api.datacite.org/events.

Reuse indicators

Data citations are the most commonly accepted metric for understanding scholarly data reuse, though usage statistics and altmetrics have also been suggested as solutions (Fear, 2013). The model of transitive credit, which describes how to measure the extent of reuse of data sets using existing standards for digital provenance, has unfortunately not yet been implemented at scale (Missier, 2016).

Reuse metrics can help evaluators understand:

- **Scholarly influence** for research datasets, both for dataset users and those that they influence (Piwowar & Vision, 2013)
- **Educational use** of datasets (Fear, 2013)
- The **diversity of contexts** in which data is reused (Fear, 2013)

The Data Citation Index is the most comprehensive and user-friendly source of citation information for research data. Using the DCI, evaluators can retrieve citation statistics for datasets and data papers that are cited by other research teams.

Citations to data can also be found in abstracting and indexing services like Dimensions, which reports over 176,000 citations to more than 22,000 Figshare-hosted datasets as of March 2020.[11]

DataCite's Event Data API is a source for other reuse metrics that more technologically adept users might consider. As of February 2020, the API reports over 107,000 "events" where a dataset has been identified as the source or derivative of another dataset. However, the technical mastery required to retrieve data via the API would be a barrier for the average evaluator.

Certain kinds of platform-specific altmetrics may be useful reuse indicators. Software sharing site Github includes metrics that allow users to measure educational reuse and the diversity of contexts in which data may be reused. The platform's native "fork" feature (where users can copy software and data for their own reuse and adaptation) provides reuse metrics. Moreover, reporting for a project's "contributors" can potentially be traced to find who has reused the data (as users who suggest revisions to code and data can be credited as project contributors on the platform).

While the vast majority of content shared on Github is software and not necessarily research related, there are high-profile examples of research data and related software being shared and reused via Github. For example, coronavirus data shared on Github[12] (itself repurposed from data originally shared by the GISAID Initiative) has been repurposed into data visualizations,[13] data analysis packages in the computing language R,[14] and training tools for data scientists.[15]

Perhaps the biggest limitation to the available reuse metrics is the uncertainty in whether these data actually point to substantive data reuse, rather than mere interest in the data or perfunctory citations for related (but not dependent) research. Though the above suggested indicators are more precise proxies for reuse, they are not direct measurements of reuse, nor of the value of reuse. For example, it is impossible to tell by looking at the numbers whether a "forked" coronavirus dataset on Github has resulted in breakthrough treatments for the illness—that is, whether the reuse of the data has resulted in so-called "real world" impacts.

This challenge is not limited to data reuse indicators; it is shared by all research indicators, including citation-based indicators. Indeed, the grand challenge of research evaluation is that all research indicators are mere proxies for impact, and not direct measures (though the terms "metrics", "measures", and "indicators" are often used interchangeably).

How to use research indicators to understand data's impact

Despite the challenges described above, evaluators can use research indicators to better understand the impact and quality of research data. Expert advice suggests that indicators be used to supplement, not supplant, expert peer review (Hicks et al., 2015; Wilsdon et al., 2015). In this section, we describe the current state of research data evaluation practices and suggest ways that the responsible use of research indicators could augment these practices.

Current use of research data indicators in evaluation

The use of quality and impact indicators in research data evaluation scenarios does not appear to be widespread. High-profile, regional initiatives such as the EU Open Science Monitor[16] and the 2014 UK Research Excellence Framework (REF) include research data in their guidelines and reporting (typically to report upon perceptions of data sharing or

[11] Statistics retrieved from app.dimensions.ai <https://app.dimensions.ai/discover/publication?search_text%3D10.6084%2520OR%252010.5061%2520OR%252010.5281%26search_type%3Dkws%26full_search%3Dtrue&sa=D&ust=1571131935649000&usg=AFQjCNE6KKqOPBecLh5DhHb1Vh7wBALakg&search_text=10.6084&search_type=kws&search_field=full_search>.

[12] https://github.com/nextstrain/ncov.

[13] https://nextstrain.org/ncov.

[14] https://github.com/GuangchuangYu/nCov2019.

[15] https://towardsdatascience.com/an-r-package-to-explore-the-novel-coronavirus-590055738ad6.

[16] https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data_en.

availability of open data repositories), but do not use indicators to assess the quality of the data itself. In fact, of the more than 190,000 research outputs submitted to the 2014 REF, only 69 were research data or databases.[17]

Indeed, this is a common theme to research data and evaluation-related policy among funding agencies, scholarly societies, and universities: while data is increasingly being acknowledged for its importance as the building block to high-quality research, data-related measures are typically concerned with the "openness" or availability of the data, rather than measuring the data's quality, impact, or reuse.

Though precedents are lacking for indicator usage, peer review guidelines (such as those summarized by Mayernik et al. (2014)) may be a useful basis for those wishing to develop methods for understanding the quality of datasets. One can also look to more general guidelines for the use of indicators in research evaluation to develop heuristics to guide how to use indicators for assessing research data.

Guidelines for using indicators to evaluate data

The Leiden Manifesto (Hicks et al., 2015) suggests a number of ways in which research indicators should be used to responsibly measure research impact. Many of these can reasonably be extended to data evaluation practices.

"Quantitative evaluation should support qualitative, expert assessment"

In cases where data is the primary focus of an evaluation, the data should be evaluated by subject area experts who are equipped to assess the data's quality and impact on its own merits. Indicators can play a part in assessment activities but should not take the place of expert review.

"Measure performance against the research missions of the institution, group or researcher"

The objectives of your organization should guide the indicators you use when assessing data's research impact, and not the other way around. Too often, evaluations are subject to the "streetlight effect" (Molas Gallart & Rafols, 2018), whereby evaluation practices develop because of the data available. Instead, you should aim to develop evaluation standards first, and then find indicators that can accurately measure your progress towards those goals.

"Account for variation by field in publication and citation practices"

When comparing data from multiple disciplines, keep in mind that each field has its own data citation and data sharing norms, and as such average citation rates may vary between fields. When setting assessment practices, it is important to document these differences in data citation and sharing norms and provide guidance for evaluators on how to compare and interpret the data.

"Scrutinize indicators regularly and update them"

The technologies used to share and cite data are ever-changing and changing rapidly. With these new technologies may come new indicators. Evaluators should periodically survey the data sharing, data citation, and altmetrics landscapes to determine if these new indicators are useful to their own assessment practices. If so, data assessment guidelines should be updated regularly to reflect best practices for using these new indicators.

"Keep data collection and analytical processes open, transparent and simple"

The Leiden Manifesto primarily advocates that metrics providers should be transparent in their data collection and aggregation processes. There is also a need for those who use data metrics to explain their impact to use metrics that are auditable and transparent. Evaluators can help researchers by writing guidelines that clearly require metric auditability and transparency.

For example, in reporting scenarios, self-collected data should be clearly explained and shared: What date did you collect these metrics? What databases did you use to collect the metrics? If you are reporting specialized or opaque indicators (e.g. the Altmetric Attention Score), have you linked to documentation that clearly explains how the indicator is calculated?

"Understand the strengths and limitations of the data"

In addition to the Leiden Manifesto's guidelines, a final important heuristic is to know the strengths and limitations of your data sources. As explained above, seemingly similar data sources (e.g. altmetrics aggregators, citation databases, etc) may collect data differently from the same source, aggregate the data according to in-house preferences, or may incorporate entirely different data sources altogether. Additionally, platform specificities like user interface design, platform searchability, or search engine optimization may affect download rates, citation practices, and so on.

Summary

In this article, we discussed how and why evaluators can use indicators to better understand the quality and impacts of research data. We provided a critical overview of the known scientometric research on research data impact indicators, including quality metrics, altmetrics, citations, usage statistics, and reuse indicators.

[17] https://results.ref.ac.uk/(S(mmua3pebw4j21yo44ae0voy2))/DownloadSubmissions/ByForm/REF2.

Current data assessment practices are primarily concerned with the openness of data, rather than with measuring quality or impact. This is an area with few precedents. We suggest that the Leiden Manifesto, originally developed for evaluation practices based on publications and their citations, may guide evaluators who seek to develop in-house heuristics for evaluating research data.

Ultimately, research data can be assessed using research indicators, like any scholarly work. Those considering adopting indicators to understand the quality and impact of research data should start from a place of expert peer review, using indicators to supplement their interpretation of the importance of a work.

Acknowledgements

The author wishes to thank Christine Burgess, Dan Valen, Mike Taylor, and Stephanie Faulkner for their input on this manuscript's early drafts. The author also thanks the anonymous reviewers for their helpful feedback.

Competing Interests

The author is employed by Altmetric.

Author Contributions

SK is solely responsible for the contents of this article.

References

Adie, E. (2013). *Gaming altmetrics*. http://www.altmetric.com/blog/gaming-altmetrics/.

Altman, M., Borgman, C., Crosas, M., & **Matone, M.** (2015). An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology*, *41*(3), 43–45. DOI: https://doi.org/10.1002/bult.2015.1720410313

Anderson, C. (2008, June 23). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. *Wired*. https://www.wired.com/2008/06/pb-theory/

Batini, C., Cappiello, C., Francalanci, C., & **Maurino, A.** (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, *41*(3), 1–52. DOI: https://doi.org/10.1145/1541880.1541883

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

Bierer, B. E., Crosas, M., & **Pierce, H. H.** (2017). Data Authorship as an Incentive to Data Sharing. *New England Journal of Medicine*, *376*(17), 1684–1687. DOI: https://doi.org/10.1056/NEJMsb1616595

Borgman, C. L. (2012). Why Are the Attribution and Citation of Scientific Data Important? In P. F. Uhlir, Board on Research Data and Information, Policy and Global Affairs, & National Research Council (Eds.), *For Attribution – Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop* (pp. 1–10). National Academies Press. http://www.nap.edu/catalog.php?record_id=13564

Bornmann, L. (2014). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, *8*(4), 895–903. DOI: https://doi.org/10.1016/j.joi.2014.09.005

Bornmann, L., & **Daniel, H.-D.** (2008). *What do citation counts measure? A review of studies on citing behavior* (Vol. 64). DOI: https://doi.org/10.1108/00220410810844150

Bornmann, L., & **Haunschild, R.** (2018). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLoS ONE*, *13*(5). DOI: https://doi.org/10.1371/journal.pone.0197133

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

Buthe, T., Jacobs, A. M., Bleich, E., Pekkanen, R., Trachtenberg, M., Cramer, K., Shih, V., Parkinson, S., Wood, E. J., Pachirat, T., Romney, D., Stewart, B., Tingley, D. H., Davison, A., Schneider, C., Wagemann, C., & **Fairfield, T.** (2015). *Transparency in Qualitative and Multi-Method Research: A Symposium* (SSRN Scholarly Paper ID 2652097). Social Science Research Network. https://papers.ssrn.com/abstract=2652097

Cabello Valdes, C., Esposito, F., Kaunismaa, E., Maas, K., McAllister, D., Metcalfe, J., O'Carroll, C., Rentier, B., Vandevelde, K., European Commission, & **Directorate-General for Research and Innovation.** (2017). *Evaluation of research careers fully acknowledging Open Science practices: Rewards, incentives and/or recognition for researchers practicing Open Science*. http://dx.publications.europa.eu/10.2777/75255

Callaghan, S., Donegan, S., Pepler, S., Thorley, M., Cunningham, N., Kirsch, P., Ault, L., Bell, P., Bowie, R., Leadbetter, A., Lowry, R., Moncoiffé, G., Harrison, K., Smith-Haddon, B., Weatherby, A., & **Wright, D.** (2012). Making Data a First Class Scientific Output: Data Citation and Publication by NERC's Environmental Data Centres. *International Journal of Digital Curation*, *7*(1), 107–113. DOI: https://doi.org/10.2218/ijdc.v7i1.218

Claibourn, M. (n.d.). *Data Types & File Formats | University of Virginia Library Research Data Services + Sciences*. Data Types & File Formats. Retrieved August 11, 2020, from https://data.library.virginia.edu/data-management/plan/format-types/

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. DOI: https://doi.org/10.2481/dsj.OSOM13-043

Costas, R., Meijer, I., Zahedi, Z., & **Wouters, P.** (2013). *The Value of Research Data: Metrics for datasets from a cultural and technical point of view* (pp. 1–48). Knowledge Exchange/Danish Agency for Culture. www.knowledge-exchange.info

COUNTER. (2014). *COUNTER | About Us.* http://www.projectcounter.org/about.html

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & **Simons, N.** (2019). Bringing Citations and Usage Metrics Together to Make Data Count. *Data Science Journal*, *18*(1), 9. DOI: https://doi.org/10.5334/dsj-2019-009

Datacite Citation Display: Unlocking Data Citations. (n.d.). [Website]. DataCite Blog. Retrieved March 3, 2020, from https://blog.datacite.org/data-citation-display/

DataCite Event Data. (n.d.). DataCite Support. Retrieved March 2, 2020, from https://support.datacite.org/docs/eventdata-guide

De Rijcke, S., Wouters, P. F., Rushforth, A. D., Franssen, T. P., & **Hammarfelt, B.** (2016). Evaluation practices and effects of indicator use-a literature review. *Research Evaluation*, *25*(2), 161–169. DOI: https://doi.org/10.1093/reseval/rvv038

Desanto, D., & **Nichols, A.** (2017). Scholarly Metrics Baseline: A Survey of Faculty Knowledge, Use, and Opinion About Scholarly Metrics. *College & Research Libraries*. DOI: https://doi.org/10.5860/crl.78.2.150

Dinsmore, A., Allen, L., & **Dolby, K.** (2014). Alternative Perspectives on Impact: The Potential of ALMs and Altmetrics to Inform Funders about Research Impact. *PLOS Biology*, *12*(11), e1002003. DOI: https://doi.org/10.1371/journal.pbio.1002003

Drachen, T., Ellegaard, O., Larsen, A., & **Dorch, S.** (2016). Sharing data increases citations. *LIBER Quarterly*, *26*(2), 67–82. DOI: https://doi.org/10.18352/lq.10149

Enkhbayar, A., Haustein, S., Barata, G., & **Alperin, J. P.** (2020). How much research shared on Facebook happens outside of public pages and groups? A comparison of public and private online activity around PLOS ONE papers. *Quantitative Science Studies*, *1*(2), 749–770. DOI: https://doi.org/10.1162/qss_a_00044

Erdt, M., Nagarajan, A., Sin, S.-C. J., & **Theng, Y.-L.** (2016). Altmetrics: An analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, 1–50. DOI: https://doi.org/10.1007/s11192-016-2077-0

Exel, M., Dias, E. L. O., & **Fruijtier, S.** (2010). *The impact of crowdsourcing on spatial data quality indicators*.

Faulkner, S. (2020, February 19). *Info on PlumX & research data* [Personal communication].

Fear, K. (2013). *The impact of data reuse: A pilot study of 5 measures*. Research Data Access & Preservation Summit, Baltimore, MD. http://www.slideshare.net/asist_org/kfear-rdap

Force11. (2015, May 11). *FORCE11 Manifesto*. FORCE11. https://www.force11.org/about/manifesto

Fox, C., Levitin, A., & **Redman, T.** (1994). The notion of data and its quality dimensions. *Information Processing & Management*, *30*(1), 9–19. DOI: https://doi.org/10.1016/0306-4573(94)90020-5

Freymann, J. B., Kirby, J. S., Perry, J. H., Clunie, D. A., & **Jaffe, C. C.** (2012). Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification. *Journal of Digital Imaging*, *25*(1), 14–24. DOI: https://doi.org/10.1007/s10278-011-9422-x

George, S. L., & **Buyse, M.** (2015). Data fraud in clinical trials. *Clinical Investigation*, *5*(2), 161–173. DOI: https://doi.org/10.4155/cli.14.116

Glänzel, W., & **Gorraiz, J.** (2015). Usage metrics versus altmetrics: Confusing terminology? *Scientometrics*, *102*(3), 2161–2164. DOI: https://doi.org/10.1007/s11192-014-1472-7

Gregg, W., Erdmann, C., Paglione, L., Schneider, J., & **Dean, C.** (2019). A literature review of scholarly communications metadata. *Research Ideas and Outcomes*, *5*, e38698. DOI: https://doi.org/10.3897/rio.5.e38698

Hahnel, M. (2013). Referencing: The reuse factor. *Nature*, *502*(7471), 298–298. DOI: https://doi.org/10.1038/502298a

Haustein, S., Bowman, T. D., Holmberg, K., Tsou, A., Sugimoto, C. R., & **Larivière, V.** (2016). Tweets as impact indicators: Examining the implications of automated "bot" accounts on Twitter. *Journal of the Association for Information Science and Technology*, *67*(1), 232–238. DOI: https://doi.org/10.1002/asi.23456

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & **Ràfols, I.** (2015). Bibliometrics: The Leiden Manifesto for Research Metrics. *Nature News*, *520*(7548), 429. DOI: https://doi.org/10.1038/520429a

Hrynaszkiewicz, I., Norton, M. L., Vickers, A. J., & **Altman, D. G.** (2010). Preparing raw clinical data for publication: Guidance for journal editors, authors, and peer reviewers. *Trials*, *11*, 9. DOI: https://doi.org/10.1186/1745-6215-11-9

Ingwersen, P., & **Chavan, V.** (2011). Indicators for the Data Usage Index (DUI): An incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics*, *12 Suppl 1*(Suppl 15), S3. DOI: https://doi.org/10.1186/1471-2105-12-S15-S3

Knepper, D., Fenske, C., Nadolny, P., Bedding, A., Gribkova, E., Polzer, J., Neumann, J., Wilson, B., Benedict, J., & **Lawton, A.** (2016). Detecting Data Quality Issues in Clinical Trials: Current Practices and Recommendations. *Therapeutic Innovation & Regulatory Science*, *50*(1), 15–21. DOI: https://doi.org/10.1177/2168479015620248

Konkiel, S. (2013). Tracking citations and altmetrics for research data: Challenges and opportunities. *Bulletin of the American Society for Information Science and Technology*, *39*(6), 27–32. DOI: https://doi.org/10.1002/bult.2013.1720390610

Konkiel, S. (2016). Altmetrics: Diversifying the understanding of influential scholarship. *Palgrave Communications*, *2*, 16057. DOI: https://doi.org/10.1057/palcomms.2016.57

Konkiel, S., & **Scherer, D.** (2013). New opportunities for repositories in the age of altmetrics. *Bulletin of the American Society for Information Science and Technology*, *39*(4), 22–26. DOI: https://doi.org/10.1002/bult.2013.1720390408

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

Kurtz, M. J., & **Bollen, J.** (2010). Usage Bibliometrics. *Annual Review of Information Science and Technology*, *44*(1), 1–64. DOI: https://doi.org/10.1002/aris.2010.1440440108

Laakso, M., & **Björk, B.-C.** (2012). Anatomy of open access publishing: A study of longitudinal development and internal structure. *BMC Medicine*, *10*(1), 124. DOI: https://doi.org/10.1186/1741-7015-10-124

Lagopoulos, A., **Tsoumakas, G.**, & **Papadopoulos, G.** (2017). Web Robot Detection in Academic Publishing. *ArXiv:1711.05098 [Cs]*. http://arxiv.org/abs/1711.05098

Lawrence, B., **Jones, C.**, **Matthews, B.**, **Pepler, S.**, & **Callaghan, S.** (2011). Citation and Peer Review of Data: Moving Towards Formal Data Publication. *International Journal of Digital Curation*, *6*(2), 4–37. DOI: https://doi.org/10.2218/ijdc.v6i2.205

Leibovici, D. G., **Rosser, J. F.**, **Hodges, C.**, **Evans, B.**, **Jackson, M. J.**, & **Higgins, C. I.** (2017). On Data Quality Assurance and Conflation Entanglement in Crowdsourcing for Environmental Studies. *ISPRS International Journal of Geo-Information*, *6*(3), 78. DOI: https://doi.org/10.3390/ijgi6030078

Leitner, F., **Bielza, C.**, **Hill, S. L.**, & **Larrañaga, P.** (2016). Data Publications Correlate with Citation Impact. *Frontiers in Neuroscience*, *10*. DOI: https://doi.org/10.3389/fnins.2016.00419

Lowenberg, D., **Chodacki, J.**, **Fenner, M.**, **Kemp, J.**, & **Jones, M. B.** (2019). *Open Data Metrics: Lighting the Fire* (Version 1) [Computer software]. Zenodo. DOI: https://doi.org/10.5281/zenodo.3525349

Mayernik, M. S., **Callaghan, S.**, **Leigh, R.**, **Tedds, J.**, & **Worley, S.** (2014). Peer Review of Datasets: When, Why, and How. *Bulletin of the American Meteorological Society*, *96*(2), 191–201. DOI: https://doi.org/10.1175/BAMS-D-13-00083.1

Meschede, C., & **Siebenlist, T.** (2018). Cross-metric compatability and inconsistencies of altmetrics. *Scientometrics*, *115*(1), 283–297. DOI: https://doi.org/10.1007/s11192-018-2674-1

Missier, P. (2016). Data trajectories: Tracking reuse of published data for transitive credit attribution. *International Journal of Digital Curation*, *11*(1), 1–16. DOI: https://doi.org/10.2218/ijdc.v11i1.425

Moed, H. F. (2016). Altmetrics as Traces of the Computerization of the Research Process. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication* (pp. 360–371). De Gruyter Saur. DOI: https://doi.org/10.1515/9783110308464-021

Molas Gallart, J., & **Rafols, I.** (2018). Why bibliometric indicators break down: Unstable parameters, incorrect models and irrelevant properties. *BiD: Textos Universitaris de Biblioteconomia i Documentació*, *40*. DOI: https://doi.org/10.2139/ssrn.3174954

Mongeon, P., **Robinson-Garcia, N.**, **Jeng, W.**, & **Costas, R.** (2017). Incorporating data sharing to the reward system of science: Linking DataCite records to authors in the Web of Science. *Aslib Journal of Information Management*, *69*(5), 545–556. DOI: https://doi.org/10.1108/AJIM-01-2017-0024

Mooney, H., & **Newton, M.** (2012). The Anatomy of a Data Citation: Discovery, Reuse, and Credit. *Journal of Librarianship and Scholarly Communication*, *1*(1), eP1035. DOI: https://doi.org/10.7710/2162-3309.1035

Mounce, R. (2013). Open access and altmetrics: Distinct but complementary. *Bulletin of the American Society for Information Science and Technology*, *39*(4), 14–17. DOI: https://doi.org/10.1002/bult.2013.1720390406

National Information Standards Organization. (2016). *Outputs of the NISO Alternative Assessment Metrics Project* (NISO RP-25-2016 Alternative Assessment Metrics Project; p. 86). National Information Standards Organization (NISO). https://www.niso.org/standards-committees/altmetrics

Nuzzolese, A. G., **Ciancarini, P.**, **Gangemi, A.**, **Peroni, S.**, **Poggi, F.**, & **Presutti, V.** (2019). Do altmetrics work for assessing research quality? *Scientometrics*, *118*(2), 539–562. DOI: https://doi.org/10.1007/s11192-018-2988-z

On the road to robust data citation. (2018). *Scientific Data*, *5*(1), 1–2. DOI: https://doi.org/10.1038/sdata.2018.95

Park, H., **You, S.**, & **Wolfram, D.** (2018). Informal data citation for data sharing and reuse is more common than formal data citation in biomedical fields. *Journal of the Association for Information Science and Technology*, *69*(11), 1346–1354. DOI: https://doi.org/10.1002/asi.24049

Pepe, A., **Goodman, A.**, **Muench, A.**, **Crosas, M.**, & **Erdmann, C.** (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. *PLOS ONE*, *9*(8), e104798. DOI: https://doi.org/10.1371/journal.pone.0104798

Peters, I., **Kraker, P.**, **Lex, E.**, **Gumpenberger, C.**, & **Gorraiz, J.** (2016). Research data explored: An extended analysis of citations and altmetrics. *Scientometrics*, *107*(2), 723–744. DOI: https://doi.org/10.1007/s11192-016-1887-4

Peters, I., **Kraker, P.**, **Lex, E.**, **Gumpenberger, C.**, & **Gorraiz, J. I.** (2017). Zenodo in the Spotlight of Traditional and New Metrics. *Frontiers in Research Metrics and Analytics*, *2*. DOI: https://doi.org/10.3389/frma.2017.00013

Pipino, L. L., Lee, Y. W., & **Wang, R. Y.** (2002). Data quality assessment. *Communications of the ACM*, *45*(4), 211–218. DOI: https://doi.org/10.1145/505248.506010

Piwowar, H. (2013). Value all research products. *Nature*, *493*(7431), 159–159. DOI: https://doi.org/10.1038/493159a

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & **Haustein, S.** (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, *6*, e4375. DOI: https://doi.org/10.7717/peerj.4375

Piwowar, H., & **Vision, T. J.** (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. DOI: https://doi.org/10.7717/peerj.175

Project COUNTER. (n.d.-a). 7.0 Processing Rules for Underlying COUNTER Reporting Data. *Project Counter*. Retrieved March 2, 2020, from https://www.projectcounter.org/code-of-practice-five-sections/7-processing-rules-underlying-counter-reporting-data/

Project COUNTER. (n.d.-b). Code of Practice for Research Data. *Project Counter*. Retrieved May 11, 2020, from https://www.projectcounter.org/code-practice-research-data/

Project COUNTER. (2018, September 13). COUNTER Code of Practice for Research Data Usage Metrics release 1. *Project Counter*. https://www.projectcounter.org/counter-code-practice-research-data-usage-metrics-release-1/

Rivalle, G., & **Green, B.** (2018). *Data Citation Index – Descriptive Document* (pp. 1–18). Clarivate Analytics. https://clarivate.libguides.com/ld.php?content_id=45722564

Robinson-Garcia, N., Jiménez-Contreras, E., & **Torres-Salinas, D.** (2015). Analyzing data citation practices according to the Data Citation Index. *Journal of the Association for Information Science and Technology*. DOI: https://doi.org/10.1002/asi.23529

Robinson-Garcia, N., Mongeon, P., Jeng, W., & **Costas, R.** (2017). DataCite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, *11*(3), 841–854. DOI: https://doi.org/10.1016/j.joi.2017.07.003

Roemer, R. C., & **Borchardt, R.** (2015). Issues, controversies, and opportunities for altmetrics. *Library Technology Reports*, *51*(5), 20-30.

Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & **Mustapha, A.** (2012). Data quality: A survey of data quality dimensions. *2012 International Conference on Information Retrieval & Knowledge Management*, 300–304. DOI: https://doi.org/10.1109/InfRKM.2012.6204995

Silvello, G. (2018). Theory and practice of data citation. *Journal of the Association for Information Science and Technology*, *69*(1), 6–20. DOI: https://doi.org/10.1002/asi.23917

Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., Haak, L. L., Haendel, M., Herman, I., Hodson, S., Hourclé, J., Kratz, J. E., Lin, J., Nielsen, L. H., Nurnberger, A., Proell, S., Rauber, A., Sacchi, S., Smith, A., ... Clark, T. (2015). Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, *1*, e1. DOI: https://doi.org/10.7717/peerj-cs.1

Stausberg, J., Bauer, U., Nasseh, D., Pritzkuleit, R., Schmidt, C. O., Schrader, T., & **Nonnemacher, M.** (2019). Indicators of data quality: Review and requirements from the perspective of networked medical research. *GMS Medizinische Informatik, Biometrie Und Epidemiologie*, *15*(1), Doc05. DOI: https://doi.org/10.3205/mibe000199

Sud, P., & **Thelwall, M.** (2014). Evaluating Altmetrics. *Scientometrics*, *98*(2), 1131–1143. DOI: https://doi.org/10.1007/s11192-013-1117-2

Sugimoto, C. (2015). 'Attention is not impact' and other challenges for altmetrics. [blog post] Wiley Exchange, June, 24, 2015.

Thelwall, M., Haustein, S., Lariviere, V., Sugimoto, C. R., Larivière, V., & **Sugimoto, C. R.** (2013). Do altmetrics work? Twitter and ten other social web. *PLOS ONE*, *8*(5), e64841. DOI: https://doi.org/10.1371/journal.pone.0064841

Wand, Y., & **Wang, R. Y.** (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, *39*(11), 86–95. DOI: https://doi.org/10.1145/240455.240479

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., Jones, R., Kain, R., Kerridge, S., Thelwall, M., Tinkler, J., Viney, I., Wouters, P., Hill, J., & **Johnson, B.** (2015). *Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management* (p. 163). HEFCE. DOI: https://doi.org/10.4135/9781473978782

Xia, J. (2012). Metrics to Measure Open Geospatial Data Quality. *Issues in Science and Technology Librarianship*, Winter. DOI: https://doi.org/10.5062/F4B85627

Zahedi, Z., Fenner, M., & **Costas, R.** (2015). *How consistent are altmetrics providers? Study of 1000 PLOS ONE publications using the PLOS ALM, Mendeley and Altmetric.com APIs.* DOI: https://doi.org/10.6084/m9.figshare.1041821.v2

Zhao, M., Yan, E., & **Li, K.** (2018). Data set mentions and citations: A content analysis of full-text publications. *Journal of the Association for Information Science and Technology*, *69*(1), 32–46. DOI: https://doi.org/10.1002/asi.23919

How to cite this article: Konkiel, S. (2020). Assessing the Impact and Quality of Research Data Using Altmetrics and Other Indicators. *Scholarly Assessment Reports*, 2(1): 13. DOI: https://doi.org/10.29024/sar.13

Submitted: 13 March 2020 **Accepted:** 09 September 2020 **Published:** 29 September 2020

Levy Library Press

Scholarly Assessment Reports is a peer-reviewed open access journal published by Levy Library Press. **OPEN ACCESS**