CONNECTING RESEARCH,
ADVANCING KNOWLEDGE

Make Data Count

Iratxe Puebla
Director, Make Data Count
GREI Metrics Webinar
October 13, 2023

@datacite
@makedatacount

@makedatacount@openbiblio.social

Data sharing is valuable, but do we understand the value of data sharing?

To nurture and incentivize data sharing we need to assign value to it

We need to understand how data are found, accessed, analyzed and utilized as part of policy development and research activities

- Who uses data & for what purposes?
- What is the impact of open data, for policy making, scientific discovery and societal benefit?
- What is the return on investment on open data?

Understanding the impact of open data requires transparent and responsible data metrics

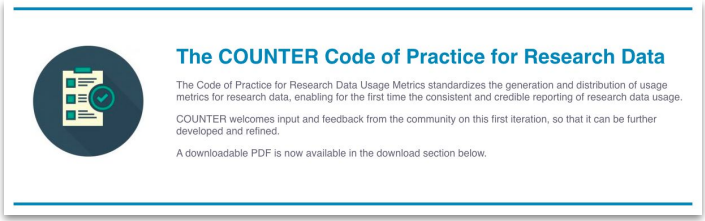
Make Data Count

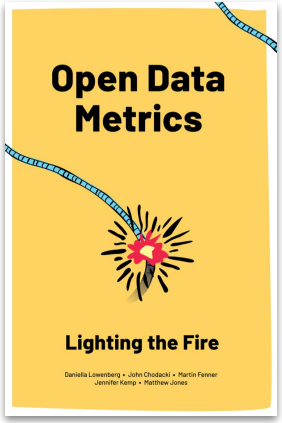
Make Data Count is an initiative that **promotes open data metrics** to enable **evaluation and reward of research data** usage and impact.

Community effort to ensure that data are used and cited in open, transparent, and responsible ways.

- **Build** open infrastructure and community-based standards.
- **Advocate** through outreach and adoption campaigns.
- **Contextualize** with evidence-based bibliometric studies.

makedatacount.org/about-us/



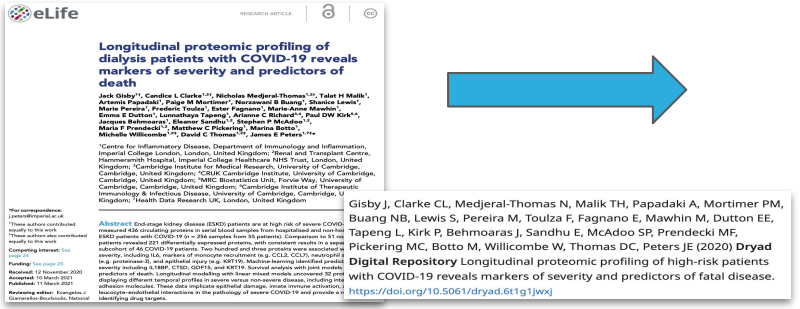


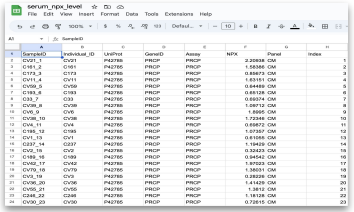
Data metrics

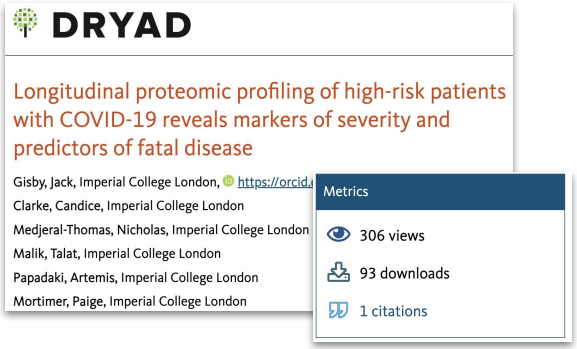
Data metrics are meaningful and contextualized quantitative or qualitative measures of how open datasets are accessed or utilized. We can collect information on data usage via:

- Views e.g. metadata, 3D models, images displayed on the landing page
- Downloads, file level or dataset level
- Citations, references to data, in the same way researchers provide a bibliographic reference to other scholarly resources

The above does not capture all possible uses of data, but they are measures that a researcher found the dataset relevant to their work in some way (views & downloads) or that the dataset has been used or reused in research (citations).







Data Metrics is a Journey.

Where do we want to get to?

Lack of incentives

Research assessment frameworks often focus on publications and do not include data

As a result, researchers perceive data sharing as a burden that brings little professional benefit

> “I don't have time. [...] It's not incentivized either [...] so why would I spend some time on something where it really amounts to no recognition [...]. People who are in this community might think I'm a good person [...] and that's fine. I can accept that. But I'm not going to track it, necessarily, because it doesn't amount to anything that's recognized.”
>
> *Professor, Anesthesiology, Canada*

Interviews by Stefanie Haustein & colleagues. Make Data Count Summit presentations. Make Data Count Summit, Washington DC. Zenodo. https://doi.org/10.5281/zenodo.8370593

Data Metrics is a Journey.

Where do we start?

Best practices for data citation

FORCE11 Data Citation Principles

FORCE11

DC1 Data Citation Principles

The Data Citation Principles cover purpose, function and attributes of citations. These principles recognize the dual necessity of creating citation practices that are both human understandable and machine-actionable.

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

1. Importance

Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications[1].

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[2].

3. Evidence

In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited[3].

4. Unique Identification

A data citation should include a persistent method for identification that is machine actionable, globally unique, and widely used by a community[4].

5. Access

Data citations should facilitate access to the data themselves and to such associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data[5].

6. Persistence

Unique identifiers, and metadata describing the data, and its disposition, should persist — even beyond the lifespan of the data they describe[6].

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[7].

8. Interoperability and Flexibility

Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that they compromise interoperability of data citation practices across communities[8].

Best practices for data citation

Guidance for publishers and data repositories

Open Access | Published: 20 November 2018

A data citation roadmap for scientific publishers

Helena Cousijn, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, Patrick Polischuk, Simone Taylor, Maryann Martone & Tim Clark

Scientific Data **5**, Article number: 180259 (2018) | Cite this article

27k Accesses | **69** Citations | **126** Altmetric | Metrics

Abstract

This article presents a practical roadmap for scholarly publishers
accordance with the Joint Declaration of Data Citation Principles
harmonization of the recommendations of major science policy b
the Publishers Early Adopters Expert Group as part of the Data C
(DCIP) project, an initiative of FORCE11.org and the NIH BioCADD
the roadmap presented here follows the "life of a paper" workflo
categories Pre-submission, Submission, Production, and Publica
intended to be publisher-agnostic so that all publishers can use
implementing JDDCP-compliant data citation. Authors reading th
know what to expect from publishers and how to enable their ow
maximum impact, as well as complying with what will become in
mandates on data transparency.

Comment | Open access | Published: 26 September 2023

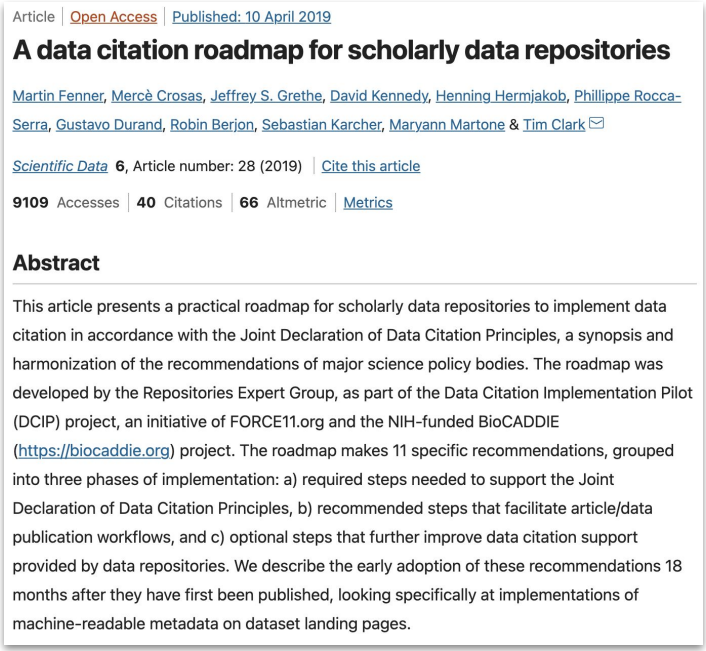
Journal Production Guidance for Software and Data Citations

Shelley Stall, Geoffrey Bilder, Matthew Cannon, Neil Chue Hong, Scott Edmunds, Christopher C. Erdmann, Michael Evans, Rosemary Farmer, Patricia Feeney, Michael Friedman, Matthew Giampoala, R. Brooks Hanson, Melissa Harrison, Dimitris Karaiskos, Daniel S. Katz, Viviana Letizia, Vincent Lizzi, Catriona MacCallum, August Muench, Kate Perry, Howard Ratner, Uwe Schindler, Brian Sedora, Martina Stockhause, ... Timothy Clark + Show authors

Scientific Data **10**, Article number: 656 (2023) | Cite this article

1573 Accesses | **29** Altmetric | Metrics

Software and data citation are emerging best practices in scholarly communication. This article provides structured guidance to the academic publishing community on how to implement software and data citation in publishing workflows. These best practices support the verifiability and reproducibility of academic and scientific results, sharing and reuse of valuable data and software tools, and attribution to the creators of the software and data. While data citation is increasingly well-established, software citation is rapidly maturing. Software is now recognized as a key research result and resource, requiring the same level of transparency, accessibility, and disclosure as data. Software and data that support academic or scientific results



Article | Open Access | Published: 10 April 2019

A data citation roadmap for scholarly data repositories

Martin Fenner, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, Robin Berjon, Sebastian Karcher, Maryann Martone & Tim Clark

Scientific Data **6**, Article number: 28 (2019) | Cite this article

9109 Accesses | **40** Citations | **66** Altmetric | Metrics

Abstract

This article presents a practical roadmap for scholarly data repositories to implement data citation in accordance with the Joint Declaration of Data Citation Principles, a synopsis and harmonization of the recommendations of major science policy bodies. The roadmap was developed by the Repositories Expert Group, as part of the Data Citation Implementation Pilot (DCIP) project, an initiative of FORCE11.org and the NIH-funded BioCADDIE (https://biocaddie.org) project. The roadmap makes 11 specific recommendations, grouped into three phases of implementation: a) required steps needed to support the Joint Declaration of Data Citation Principles, b) recommended steps that facilitate article/data publication workflows, and c) optional steps that further improve data citation support provided by data repositories. We describe the early adoption of these recommendations 18 months after they have first been published, looking specifically at implementations of machine-readable metadata on dataset landing pages.

www.nature.com/articles/sdata2018259
www.nature.com/articles/s41597-023-02491-7

www.nature.com/articles/s41597-019-0031-8

Data usage initiatives

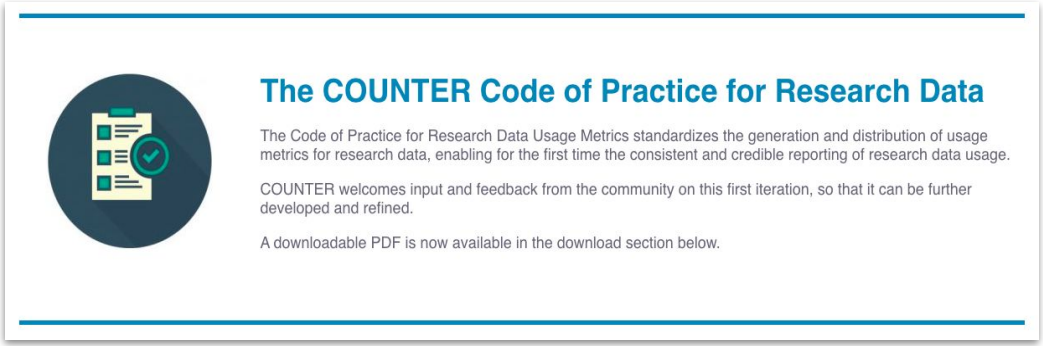
COUNTER Code of Practice for Research Data

Provides a standard for data repositories and platform providers to standardize the generation and distribution of usage metrics for research data

Initially released in 2018

Implemented by Dryad, Figshare, Zenodo, Vivli and others

Planning for next version of the Code



www.projectcounter.org/code-of-practice-rd-sections/foreword/

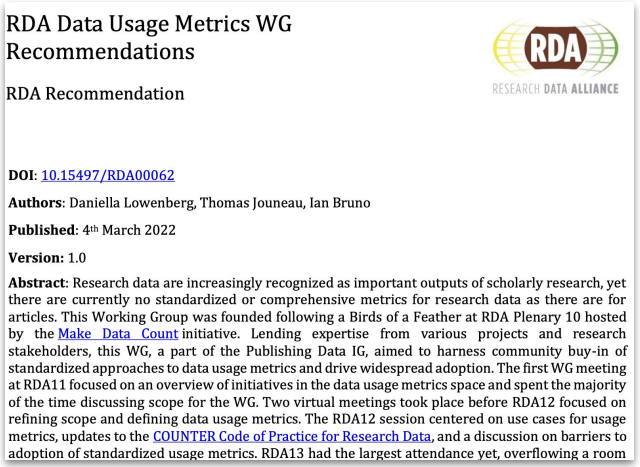
Data usage initiatives

RDA Data Usage Metrics Working Group

All stakeholders will benefit from metrics that can reliably reflect impact of research data, for this to happen, data usage at repositories should be reliably counted and reported

Recommendations:

- Strong interest in how to meaningfully report data views & downloads, but there are activation barriers ⇒ Keep **focus on adoption of minimal frameworks**
- **Endorsed the COUNTER Code of Practice** for reporting views & downloads, and utilizing **DataCite for aggregations**
- Data usage is nuanced, we **should not default to a 'data impact factor'**



RDA Data Usage Metrics WG Recommendations

RDA Recommendation

DOI: 10.15497/RDA00062

Authors: Daniella Lowenberg, Thomas Jouneau, Ian Bruno

Published: 4th March 2022

Version: 1.0

Abstract: Research data are increasingly recognized as important outputs of scholarly research, yet there are currently no standardized or comprehensive metrics for research data as there are for articles. This Working Group was founded following a Birds of a Feather at RDA Plenary 10 hosted by the Make Data Count initiative. Lending expertise from various projects and research stakeholders, this WG, a part of the Publishing Data IG, aimed to harness community buy-in of standardized approaches to data usage metrics and drive widespread adoption. The first WG meeting at RDA11 focused on an overview of initiatives in the data usage metrics space and spent the majority of the time discussing scope for the WG. Two virtual meetings took place before RDA12 focused on refining scope and defining data usage metrics. The RDA12 session centered on use cases for usage metrics, updates to the COUNTER Code of Practice for Research Data, and a discussion on barriers to adoption of standardized usage metrics. RDA13 had the largest attendance yet, overflowing a room

10.15497/RDA00062

Data Metrics is a Journey.

Where are we now?

Data usage: where are we?

The COUNTER Code of Practice has not yet been implemented as broadly as we would like

- Processes for normalizing data usage can be time consuming for repositories
- Requires developer understanding of the Code of Practice
- Code to maintain: log processing and SUSHI report generation

DataCite usage tracker
Facilitates implementation of data usage reporting via a tracker that collects web-based usage

Currently in beta

Collects repository usage stats using a Javascript tracker—not log file processing

DataCite generates monthly reports for repositories

HTML

```
<!-- Track View -->
<script defer data-doi="10.5072/1234"
        data-repoid="example.com"
        data-metric="view"
        src="https://cdn.jsdelivr.net/npm/@datacite/datacite-tracker"></script>

<!-- Track Download -->
<script defer data-doi="10.5072/1234"
        data-repoid="example.com"
        data-metric="download"
        src="https://cdn.jsdelivr.net/npm/@datacite/datacite-tracker"></script>
```

support.datacite.org/docs/datacite-usage-tracker

Data Citations: where are we?

Data citation workflow requires several steps involving different stakeholders in order for the information to propagate

Researchers do not always cite the data they use

Not all repositories capture citation information

Some repositories use accession numbers instead of DOIs

Citations and/or associated metadata may be lost when the publisher deposits in Crossref

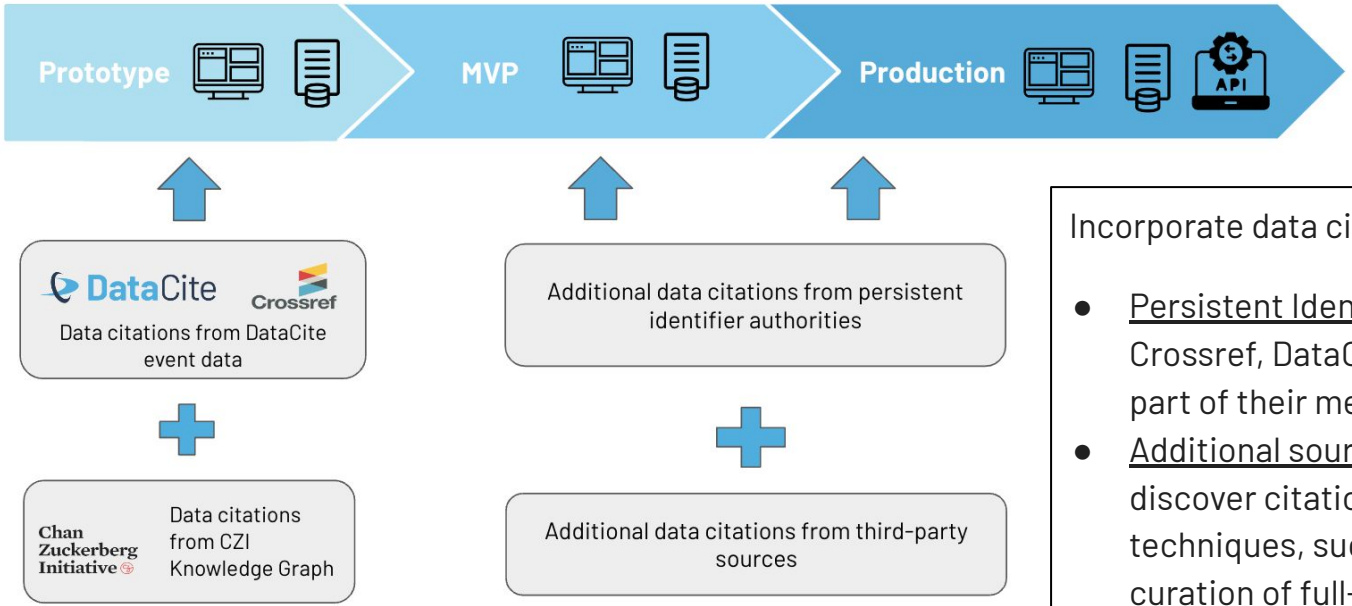
Event data

We know there are many more instances of data usage than we are currently capturing

Workflow adapted from Susan Borda 'If Data is Used in the Forest and No-one is Around to Hear it, Did it Happen? a Citation Count Investigation'.
Traffic light icon by Freepik via Flaticon

Global Open Data Citation Corpus

Goal: **Develop a comprehensive corpus that incorporates data citations from different sources into a centralized, publicly accessible community resource**



Incorporate data citations from diverse sources:

- Persistent Identifier (PID) authorities (e.g. Crossref, DataCite) that collect citations as part of their metadata deposit workflow.
- Additional sources that aggregate or discover citations through various techniques, such as machine learning and curation of full-text in articles.

Data Citation Corpus: Prototype

We are finalizing the prototype for the corpus, which will include data citations from DataCite event data and data mentions from CZI Science Knowledge Graph.

- Basic user interface to visualize the data with different filters
- Seed data available via a data dump

makedatacount.org/data-citation/
http://corpus.stage.datacite.org/dashboard

The time is now: Make Data Count Summit

Meeting in Washington DC, September 2023 - Brought together representatives from research institutions, funders and government, researchers, publishers, and infrastructure providers for focused discussions on data evaluation & data metrics

Takeaways:

- Standards and infrastructure are available to enable adoption, we need to support those and **SCALE** the data usage information available to the community
- **Assessment frameworks** at institutions need to incorporate evaluations of data reuse, this needs engagement at all institutional levels and incremental approaches
- Data metrics must align to the principles for open data and be **complete, transparent, auditable, and contextualized.**

makedatacount.org/2023/09/28/make-data-count-summit-the-time-is-now-to-advance-data-metrics/

The time is now: Help us complete the picture

- **Researchers**: cite the datasets you use
- **Repositories:**
  - Collect data usage and citations for datasets
  - Submit data usage and citations to DataCite
- **Researchers, repositories, institutions, funders**: Provide input on corpus development

We want to hear success stories! Collaborate with Make Data Count on outreach for the repository community

iratxe.puebla@datacite.org
info@datacite.org

Make Data Count: makedatacount.org

Photo by Ryoji Iwata via Unsplash

Open Metrics Subcommittee

Mark Hahnel, Figshare
Co-Chair

Matt Buys - Datacite
Co-Chair

Jim Myers, Dataverse

C.3.5 Open Metrics Goals

Year 2 Goals

3.5.1 Roadmap for the technical priorities - Develop a multi-year roadmap in line with the priorities defined such as implementation of usage tracker, relational metadata submitted to DataCite, Priority metadata, and reverse engineering legacy usages statistics.

3.5.2 Priority metadata - Define the metadata properties needed to address the metrics use cases and coordinate with use cases sub-committee. In addition, the sub-committee will explore whether aggregators can donate seed file metadata regarding subject classification metadata.

3.5.3 Definitions - Establish clear definitions for metrics, leveraging existing standards and community initiatives. The intention is to standardize the definition of metrics across the generalist repositories.

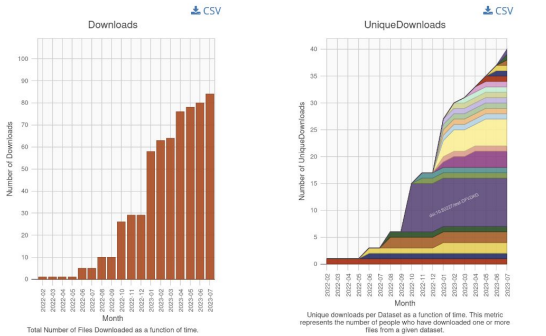
3.5.4 Baseline metrics - Develop a roadmap for baseline metrics, defining what metrics we current have available, the limitations and the timelines to begin using data metrics in research evaluation (see https://makedatacount.org/data-metrics-2/)

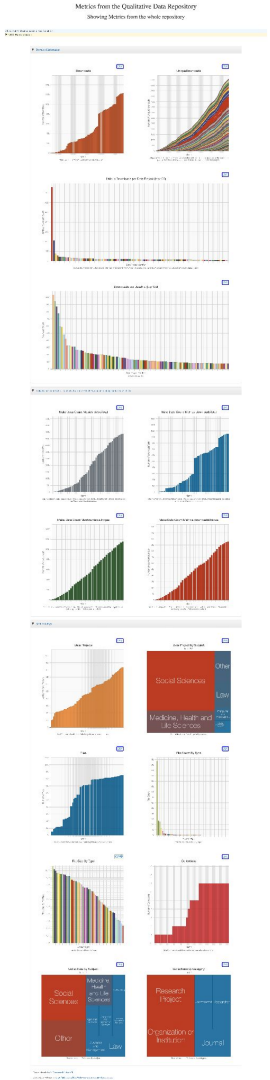
Subcommittee Progress Year 2

- Additional Open Metrics Y2 goals are 46% complete
- Fulfillment of some goals dependent on development of DataCite usage tracker, several repositories have additional work planned to get data sent to DataCite

	Implementation of Make Data Count (MDC) as standard	Development and implementation of a usage tracker	Submission of relational metadata to DataCite	Real-time repository metrics measuring usage and impact on a publicly accessible page	Monthly repository metrics report of NIH-funded research
Dataverse	In progress	In progress	NO RESPONSE	In progress	In progress
Dryad	Completed prior	Completed prior	Completed prior	Completed prior	Completed Y1
Figshare	Completed prior	Planned Y3/Y4	Completed in Y2	Planned Y2	In progress
Mendeley	Planned Y3/Y4	In progress	NO RESPONSE	Planned Y2	Planned Y3/Y4
OSF	In progress	Completed Y1	Completed in Y2	Planned Y3/Y4	In progress
Vivli	Completed Y1	Completed Y2	Planned Y3/Y4	Planned Y3/Y4	Completed Y2
Zenodo	Completed prior	Completed prior	Completed prior	Planned Y2	Planned Y2

Metrics from the Qualitative Data Repository (https://data.qdr.syr.edu/metrics_5ef2ae2be4b)





Subcommittee Work at the Repository Level

Subcommittee Progress

- The Open Metrics Subcommittee has been actively engaged in discussions and initiatives related to data metrics and usage within the academic and research community over the past two years.
- Significant progress has been made in promoting standardization, usage tracking, and better metadata practices.

Repository Metrics Roadmap

- Repositories have made substantial progress in implementing key metrics, such as COUNTER standards, views and downloads tracking, and collecting author-related articles.
- Progress varies across different repositories, with some already completing several milestones.

Examples of being able to track NIH programmatic research are as follows:

- All DataCite DOIs with NIH in the funderIdentifier field (either a Crossref Funder ID or ROR ID): https://api.datacite.org/dois?query=fundingReferences.funderIdentifier%3A%22https%3A%2F%2Fdoi.org%2F10.13039%2F100000002%22+OR+fundingReferences.funderIdentifier%3A%22https%3A%2F%2Fror.org%2F01cwqze88%22&registration-agency=datacite

- Filtered by repository: https://api.datacite.org/dois?client-id=dryad.dryad&query=fundingReferences.funderIdentifier:%22https://doi.org/10.13039/100000002%22%20OR%20fundingReferences.funderIdentifier:%22https://ror.org/01cwqze88%22

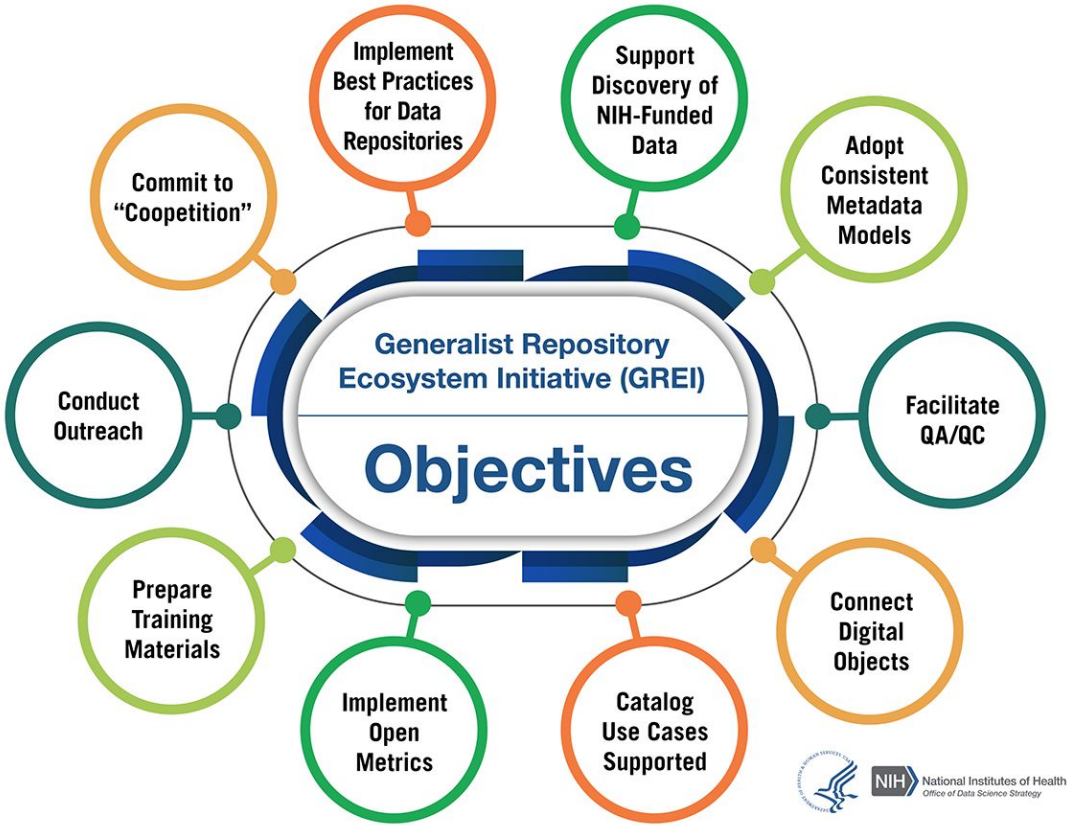
Subcommittee Impact

Collaboration and Trajectory:

- Collaboration among repositories within GREI groups is crucial for success in implementing data metrics.
- Repositories are transitioning from adopting best practices to a more nuanced approach, allowing for tailored solutions to specific use cases and downstream requirements.

NIH Programmatic Research Tracking:

- Repositories are actively tracking NIH programmatic research through DataCite DOIs with NIH in the funderIdentifier field, providing transparency and accountability.



Looking forward

Future Initiatives:

- Repositories are committed to key milestones, including adopting MDC as a standardized practice, developing a reliable usage tracker, submitting comprehensive relational metadata to DataCite, establishing real-time repository metrics, and generating monthly metrics reports for NIH-funded research.

Commitment to Data Metrics:

- The report underscores the collective commitment of repositories to advance data metrics, emphasizing comprehensive and contextualized data evaluation practices.

Next Steps:

- Continue collaboration and coordination efforts among repositories to achieve data metrics goals.
- Monitor progress on Year 2 goals and adapt strategies as needed to ensure successful implementation.
- Carry efforts into the task group structure
- Seeing how citations fit in to the puzzle - to then help drive standardization

Unpublished State of Open Data 2023