



CONNECTING RESEARCH,
ADVANCING KNOWLEDGE

6.2 A Radical New Approach to Data Citation: Cook the Carrots, Burn the Sticks

Jamie Wittenberg - University of Colorado Boulder

John Chodacki - California Digital Library

Kristi Holmes - Northwestern University

12 December 2023
CNI 2023



@makedatacount



@makedatacount@openbiblio.social



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

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

Sources Data Citations

Source	Connections
DataCite to Crossref	5,646,993
DataCite Related Identifiers	367,347
Crossref to DataCite	7,629
Crossref Related Identifiers	123
Crossref to DataCite*	101

Numbers from Event Data Service 20 April 2021

The Stick or the Carrot? An Approach to Open Access

[Randi Tyse Eriksen](#), *Norwegian University of Science and Technology*

Follow

[Jorunn Alstad](#), *Norwegian University of Science and Technology*

Follow

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Neither carrots nor sticks? Challenges surrounding data sharing from the perspective of research funding agencies —A qualitative expert interview study

Michael Anger, Christian Wendelborn, Eva C. Winkler, Christoph Schickhard

Published: September 7, 2022 • <https://doi.org/10.1371/journal.pone.0273259>

Measuring the Norm of Reciprocity on Data Sharing Practices: A Carrot or Stick Approach?

[Crystal Pleake Sherline](#), *University of Tennessee - Knoxville*

Follow

RESEARCH NOTE

Cash, carrots, and sticks: Open Access incentives for researchers [version 1; peer review: 1 approved, 1 approved with reservations]

[Joseph Kraus](#)

[Author details](#)

University of Denver, Denver, CO, 80208, USA

Commentary | [Published: 30 October 20](#)

Of carrots and sticks

[Jens Kattge](#), [Sandra Díaz](#) & [Christian V](#)

OPEN ACCESS

COMMUNITY PAGE

Sharing Neuron Data: Carrots, Sticks, and Digital Records

Giorgio A. Ascoli

Published: October 8, 2015 • <https://doi.org/10.1371/journal.pbio.1002275>

Sticks and carrots: encouraging open science at Carrots and Sticks

Sabina Leonelli, Daniel Spichtinger, Barbara Prainsack

First published: 23 March 2015 | <https://doi.org/10.1002/geo2.2> | Citations: 31

This paper was accepted for publication in November 2014

Some Ideas on How to Create a Successful Institutional Repository

[Miguel Ferreira](#)

Department of Information Systems of the University of Minho, Portugal
<mferreira@dsi.uminho.pt>

[Ana Alice Baptista](#)

Department of Information Systems of the

[Eloy Rodrigues](#)

Documentation Services of the University of Minho, Portugal
<eloy@s dum.uminho.pt>

[Ricardo Saraiva](#)

Documentation Services of the University of Minho, Portugal
<rsaraiva@s dum.uminho.pt>

Carrots and Sticks: A Qualitative Study of Library Responses to the UK's Research Excellence Framework (REF) 2021 Open Access Policy

Dan DeSanto

Research Policy as “Carrots and Sticks”: Governance Strategies in Australia, the United Kingdom and New Zealand

[Jenny M. Lewis](#)

Stick or carrot - how to fill an institutional repository

Staffan Parnell, Senior Librarian, SLU Libraries, Ultuna

[Check for updates](#)





Sources Data Citations

Source	Connections
DataCite to Crossref	5,646,993
DataCite Related Identifiers	367,347
Crossref to DataCite	7,629
Crossref Related Identifiers	123
Crossref to DataCite*	101

Numbers from Event Data Service 20 April 2021

The Global Data Citation Corpus!



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.

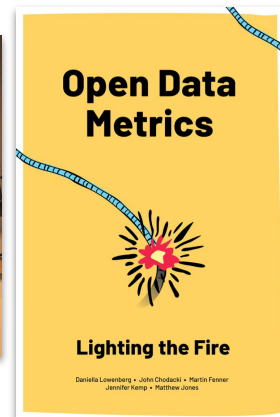
- **Build** open infrastructure and community-based standards.
- **Advocate** through outreach and adoption campaigns.
- **Contextualize** with evidence-based bibliometric studies.



The COUNTER Code of Practice for Research Data

The Code of Practice for Research Data Usage Metrics standardizes the generation and distribution of usage metrics for research data, enabling for the first time the consistent and credible reporting of research data usage. COUNTER welcomes input and feedback from the community on this first iteration, so that it can be further developed and refined.

A downloadable PDF is now available in the download section below.



makedatacount.org/about-us

Journey to “Data Metrics”

Where are we now?

Step 1

Determine
community
best practice

Step 2

**Adopt best
practices**

Step 3

Contextualize
best practices

Step 4

Use
data metrics to
enable evaluation

Step 5

Incentivise
researchers to
share data

Information on data usage

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

eLife

RESEARCH ARTICLE

Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death

Jack Gibby¹, Candice L. Clarke^{1*}, Nicholas Medjeral-Thomas^{1*}, Tahir H Malik¹, Arsenia Papadaki¹, Pooja M Mortimer¹, Hacercan B Buzug¹, Shalika Lewis¹, Maria Pereira¹, Frederic Touza¹, Ester Fagnano¹, Marie-Anne Mawhin¹, Emma E Dutton¹, Lounisahe Traang¹, Adam C Pickering¹, Paul DW Kirk¹, Jonathan Behrman¹, Eleanor Satchell¹, Sheehan P McLaughlin¹, Maria P Predecki¹, Matthew C Botto¹, Maura Botton¹, Michelle Willicombe¹, David C Thomas¹, James E Peters¹

¹Centre for Inflammatory Disease, Department of Immunology and Inflammation, Imperial College London, London, United Kingdom; ²Renal and Transplant Centre, Hammersmith Hospital, Imperial College Healthcare NHS Trust, London, United Kingdom; ³Cambridge Institute for Medical Research, University of Cambridge, Cambridge, United Kingdom; ⁴CRUK Cambridge Institute, University of Cambridge, Cambridge, United Kingdom; ⁵INSIS, Sorbonne Univ, Paris, France; ⁶University of Cambridge, Cambridge, United Kingdom; ⁷Cambridge Institute of Therapeutic Immunology & Infectious Disease, University of Cambridge, Cambridge, United Kingdom; ⁸Health Data Research UK, London, United Kingdom

*Correspondence: j.gibby@imperial.ac.uk

These authors contributed equally to this work.

These authors also contributed equally to this work.

Competing interest: See [https://doi.org/10.1371/journal.pone.0240636.g001](#)

Copyright: © 2021 Gibby et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

RESEARCH ARTICLE

Abstract End-stage kidney disease (ESKD) patients are at high risk of severe COVID-19. We measured the circulating proteome in serial blood samples from hospitalized and non-hospitalized ESKD patients with COVID-19 in a 256 samples from 50 patients. Comparison to 51 non-infected patients revealed 21 differentially expressed proteins, with enrichment results to a subpopulation of 46 COVID-19 patients. Two hundred and three proteins were associated with clinical severity, including 16 markers of acute organ impairment (e.g. CCL5, CXCL8, myeloperoxidase) (e.g. proinflamm-3), and epithelial injury (e.g. AREG). Machine-learning identified predictors of severity including IL-18IP, CXCL10, CXCL8, and AREG. Spatial analysis with laser capture revealed 69 proteins in situ. Longitudinal monitoring with three renal models identified 63 proteins displaying different temporal profiles in severe versus non-severe disease, including integrins and adhesion molecules. These data indicate epithelial damage, under immune activation, and neutrophil-mediated interactions in the pathology of severe COVID-19 and provide a resource for identifying drug targets.

Keywords: COVID-19, ESKD, proteomics, kidney disease, dialysis, COVID-19 severity, predictors of death, acute organ impairment, epithelial injury, integrins, adhesion molecules

RESEARCH ARTICLE

RESEARCH ARTICLE



Gisby J, Clarke CL, Medjeral-Thomas N, Malik TH, Papadaki A, Mortimer PM, Buang NB, Lewis S, Pereira M, Toulza F, Fagnano E, Mawhin M, Dutton EE, Tapeng L, Kirk P, Behmoaras J, Sandhu E, McAdoo SP, Predecki MF, Pickering MC, Botto M, Willicombe W, Thomas DC, Peters JE (2020) **Dryad Digital Repository** Longitudinal proteomic profiling of high-risk patients with COVID-19 reveals markers of severity and predictors of fatal disease. <https://doi.org/10.5061/dryad.6t1g1wjxj>

serum_npx_level

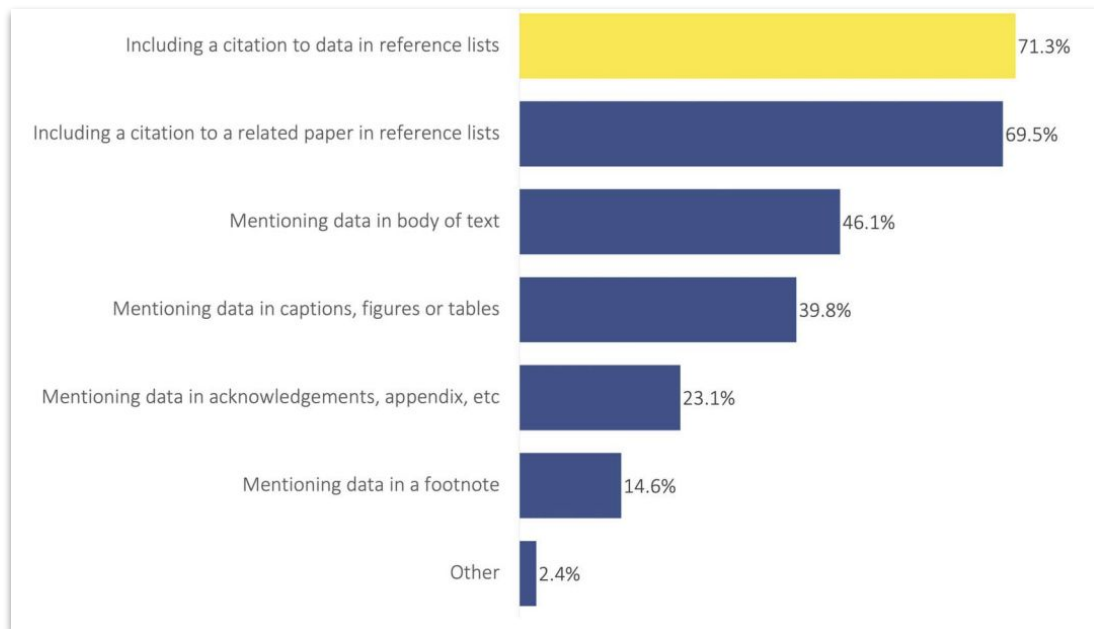
File Edit View Insert Format Data Tools Extensions Help

100% 123 Default...

A1	A	B	C	D	E	F	G	H
	SampleID	Individual_ID	UNProt	GeneID	Assay	NPX	Panel	Index
1	CV21	CV21	P42785	PRCP	PRCP	2.209598	CM	1
2	CV1_1	CV1_1	P42785	PRCP	PRCP	1.582986	CM	2
3	CV12_3	CV12_3	P42785	PRCP	PRCP	0.889872	CM	3
4	CV11_4	CV11_4	P42785	PRCP	PRCP	1.651151	CM	4
5	CV9_5	CV9_5	P42785	PRCP	PRCP	0.644646	CM	5
6	CV193_6	CV193_6	P42785	PRCP	PRCP	0.65128	CM	6
7	CV3_7	CV3_7	P42785	PRCP	PRCP	0.68374	CM	7
8	CV9_8	CV9_8	P42785	PRCP	PRCP	1.09712	CM	8
9	CV6_9	CV6_9	P42785	PRCP	PRCP	1.18995	CM	9
10	CV9_10	CV9_10	P42785	PRCP	PRCP	1.72346	CM	10
11	CV4_11	CV4_11	P42785	PRCP	PRCP	0.69972	CM	11
12	CV19_12	CV19_12	P42785	PRCP	PRCP	1.07071	CM	12
13	CV1_13	CV1_13	P42785	PRCP	PRCP	0.61055	CM	13
14	CV37_14	CV37_14	P42785	PRCP	PRCP	1.18429	CM	14
15	CV2_15	CV2_15	P42785	PRCP	PRCP	0.32423	CM	15
16	CV19_16	CV19_16	P42785	PRCP	PRCP	0.94542	CM	16
17	CV42_17	CV42_17	P42785	PRCP	PRCP	1.91023	CM	17
18	CV7_18	CV7_18	P42785	PRCP	PRCP	1.38031	CM	18
19	CV3_19	CV3_19	P42785	PRCP	PRCP	0.26226	CM	19
20	CV6_20	CV6_20	P42785	PRCP	PRCP	1.41429	CM	20
21	CV9_21	CV9_21	P42785	PRCP	PRCP	1.1812	CM	21
22	CV4_22	CV4_22	P42785	PRCP	PRCP	1.18128	CM	22
23	CV9_23	CV9_23	P42785	PRCP	PRCP	0.72615	CM	23

Data citations

- Valued by researchers
- Workflows available for repository & publishers
- AI tools developed to identify data citations



Survey on data citation practices. Preferences for how respondents would like others to refer to their own data, n = 2,454.

Looking for data citations



Photo by Matt Howard via Unsplash

Challenges: People

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

⇒ Researchers are **not incentivized to cite data**

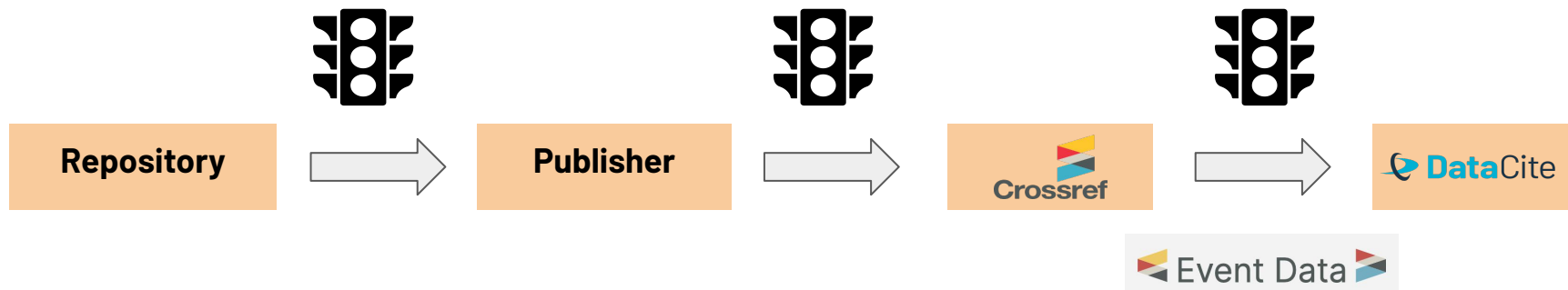
Some institutions and funders are seeking to include data as part of their processes, but lack the tools to gather data usage measures in a comprehensive, transparent and responsible manner.

“If it's done right, it's okay. If it's done, like, let's say the h-index then oh my god, please no.”

Postdoc, HCI, Germany

Challenges: Technologies

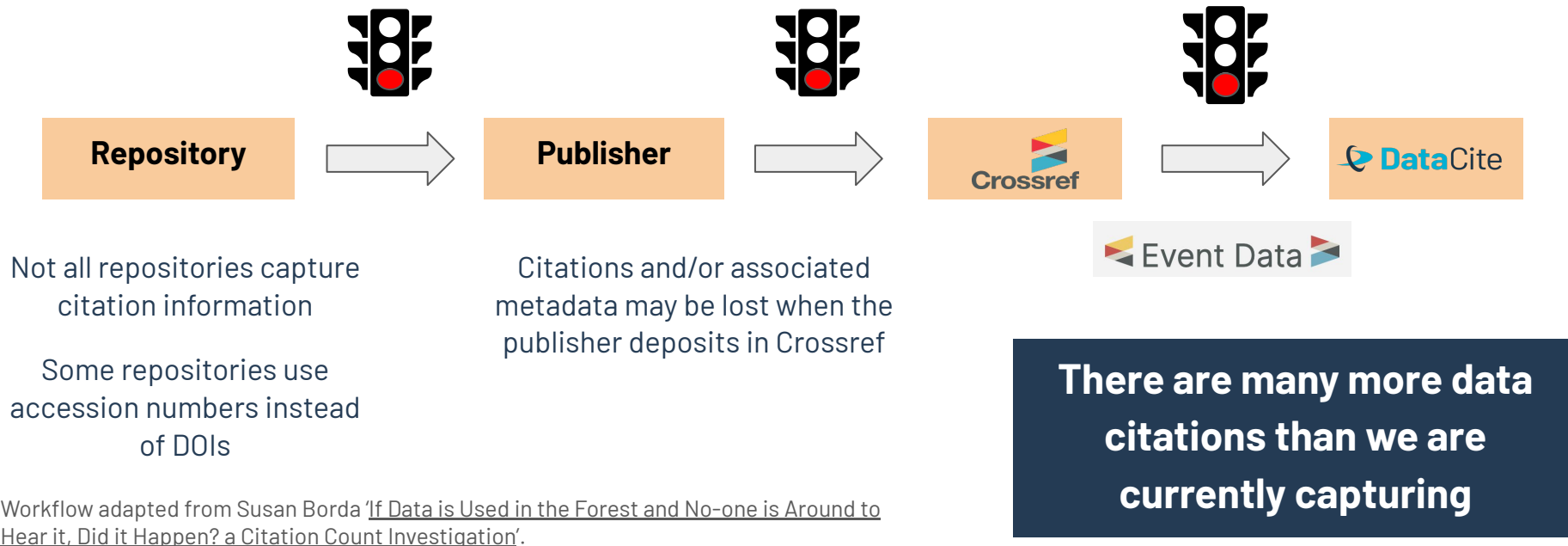
The data citation workflow requires several steps involving different stakeholders in order for the information to propagate.



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'](#).

Challenges: Technologies

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

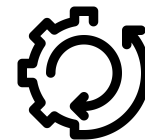


Workflow adapted from Susan Borda [‘If Data is Used in the Forest and No-one is Around to Hear it, Did it Happen? a Citation Count Investigation’](#).

Challenges: Ecosystem

Data citations can be collected via PID registration workflows, but there are also citations that do not make it to PID metadata.

A number of stakeholders have created approaches to find citations to data, but these citations are stored in different locations.



Other groups utilizing AI tools, curation etc

The community lacks a single comprehensive resource to access data citations

Addressing the challenges to enable easy discovery & use of data citations

Photo by Nick Fewings via Unsplash



Global Data Citation Corpus




A comprehensive corpus that incorporates data citations from different sources into a centralized, publicly accessible community resource

Supported by the Wellcome Trust

Collaborative effort to incorporate data citations from PID metadata deposit workflows as well as additional sources that aggregate or discover citations through various techniques, such as machine learning and curation of full-text in articles.

Collaboration with Chan Zuckerberg Initiative

Chan
Zuckerberg
Initiative 

EMBL-EBI 

- Leverage CZI machine learning model and access to a large corpus of articles to identify data mentions
- Include citations associated with **accession numbers from Europe PMC**

Global Data Citation Corpus



- Foundation for collection of citations from different sources
- Initial dashboard
- Data dump

- Collaborations to extend coverage & inform developments for different users
- Enhanced dashboard
- Data dump + API


- Coverage & features that enable use in evaluation processes & bibliometrics studies
- Enhanced dashboard
- Data dump + API



 
Data citations from DataCite event data

Additional data citations from persistent identifier authorities



 Data citations from CZI Knowledge Graph

Additional data citations from third-party sources

Global Data Citation Corpus



Prototype



MVP



Production



Data citations from DataCite event data



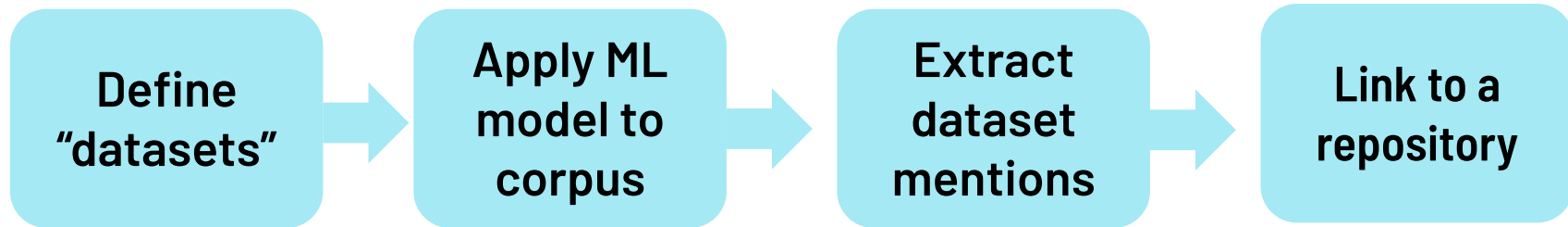
Data citations from CZI Knowledge Graph

Additional data citations from persistent identifier authorities



Additional data citations from third-party sources

Corpus prototype



A collection of data that have been measured, collected, and/or analyzed as part of a research study

- Accession Number IDs
- Repository DOIs
- Resources hosted on external URLs such as academic institutions

SciBERT-based Named Entity Recognition



CZI Full-Text, Europe PMC Full-Text



Extract dataset mentions

Extract mentions based on the persistent identifiers

Clean & disambiguate

The microarray data had been previously deposited at Gene Expression Omnibus (GEO) under accession number **GSE2603**.

Link to a repository

Join mentions with the dataset record

<https://identifiers.org/geo:GSE2603>

Global Data Citation Corpus: Prototype



We are finalizing the prototype for the corpus, which includes:

- Data citations from DataCite Event Data
- Data mentions from CZI Science Knowledge Graph



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

Prototype user interface

The Global Data Citation Corpus Community!

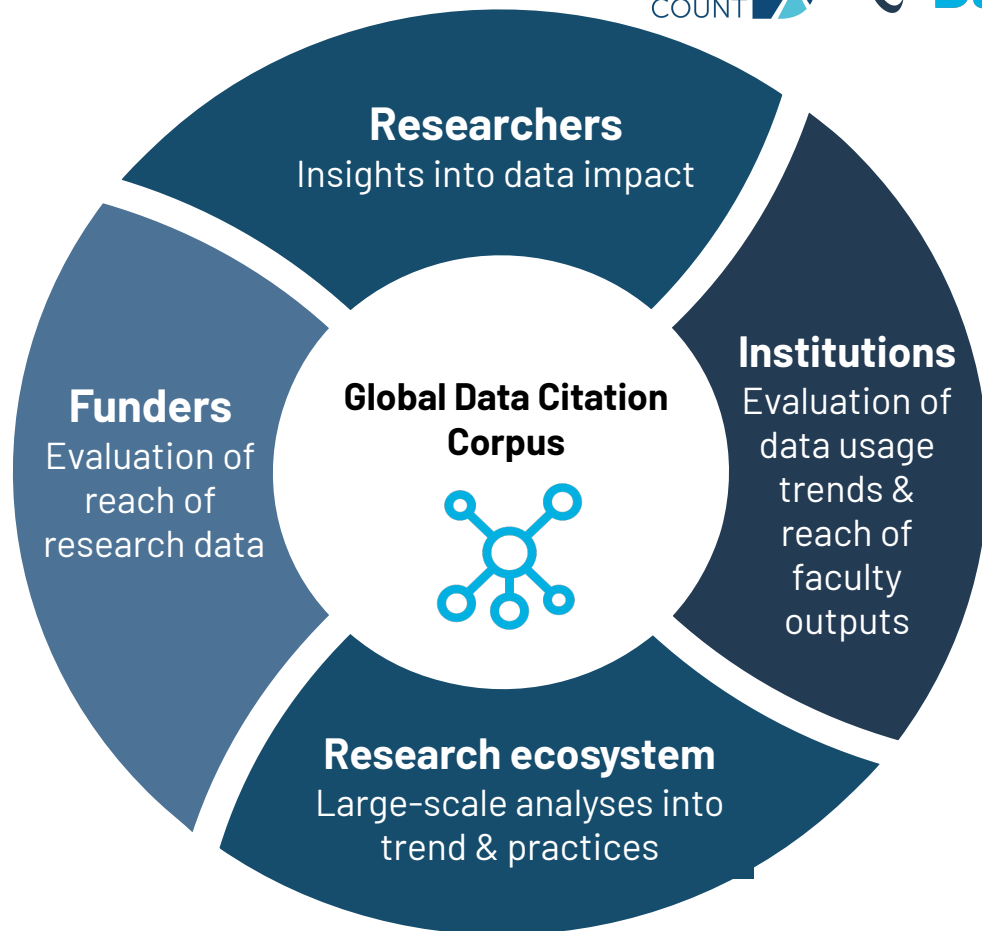


Global Data Citation Corpus



Paving the way to the evaluation of data usage

Partnering *across the community* to support open, equitable, and meaningful data and tools about data



The Community:

Building the Corpus Together



Repositories

- Submit citations through DOI metadata
- Track and display data citations
- Share feedback!



Publishers

- Include data citations in Crossref metadata deposit
- Engage with us in discussions



Institutions and Funders

- Provide feedback to inform how to best align corpus to uses in institutional and funder processes



Organizations with known Data Citations

- Submit citations to the open data citation corpus - *Interested? Do get in touch to discuss!*

**Libraries as the
connective tissue**

The Community:

Libraries as connective tissue

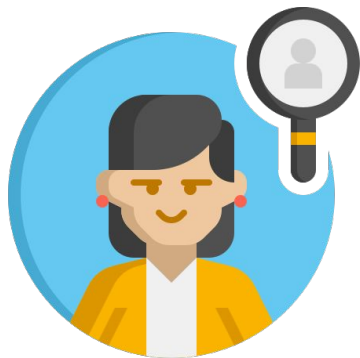


Icons by Freepik

**Libraries support understanding and action across the spectrum
- from superlocal to universal levels and topics**

The Community:

Stay informed, get involved



[Icon by Paul J.](#)

Stay informed!

- Project website: <https://makedatacount.org/data-citation/>
- Join the listserv: <https://www.tinyurl.com/makedatacountupdates>
- Recent slides: <https://zenodo.org/records/10083792>

Get involved!

- Engage with the community, provide feedback and ideas: <https://tinyurl.com/datacitationcorpus>
- Take the tools out for a test drive, e.g., <http://corpus.stage.datacite.org/dashboard>
- Attend MDC events - webinars, MDC Summit Fall 2024!



Coming in March 2024:

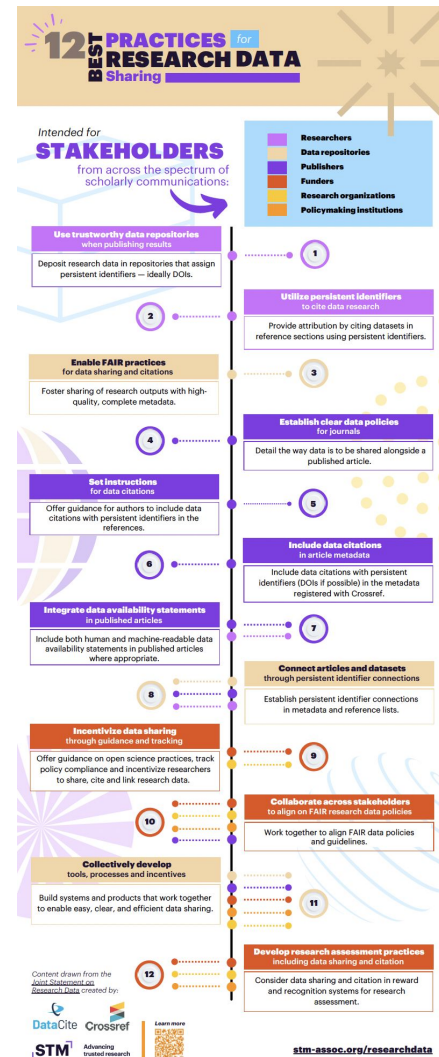
MDC paper in upcoming Harvard Data Science Review “Democratizing Data: Data as a Public Asset” special issue: “Building Trust: Data Metrics as a Focal Point for Responsible Data Stewardship”

What else? Discuss and Endorse! Joint Statement on Research Data

The Joint Statement on Research Data

- Developed by by STM, DataCite, and CrossRef
- Calls for adoption of best practices for data sharing and data citation.
- Recommendations and Infographic

Signatories adopt and promote the relevant best practices; action is intended to inspire the community, including researchers, research funders, research institutions, data repositories and publishers, to join us in making it easy for researchers to share, link and cite research data.



Thank you!

Make Data Count:

makedatacount.org

Data citation corpus:

makedatacount.org/data-citation

We will seek community input as we work on the development of the data citation corpus. Interested in learning more? Do get in touch:

iratxe.puebla@datacite.org

info@datacite.org

**Thanks to Wellcome Trust, Chan
Zuckerberg Initiative, EMBL-EBI, DataCite**

Generalist Repository Ecosystem Initiative