

Embedding data in research evaluation: Data citations unlock insights into data usage and impact

Iratxe Puebla
Director, Make Data Count

CiVers Workshop
14 October 2025





An initiative that works to build the tools and practices necessary so that the community can meaningfully assess how data are used.

Our vision: an ecosystem where **data are routinely evaluated and rewarded** as primary outputs.



Open infrastructure
to collect and share
measures of data
usage

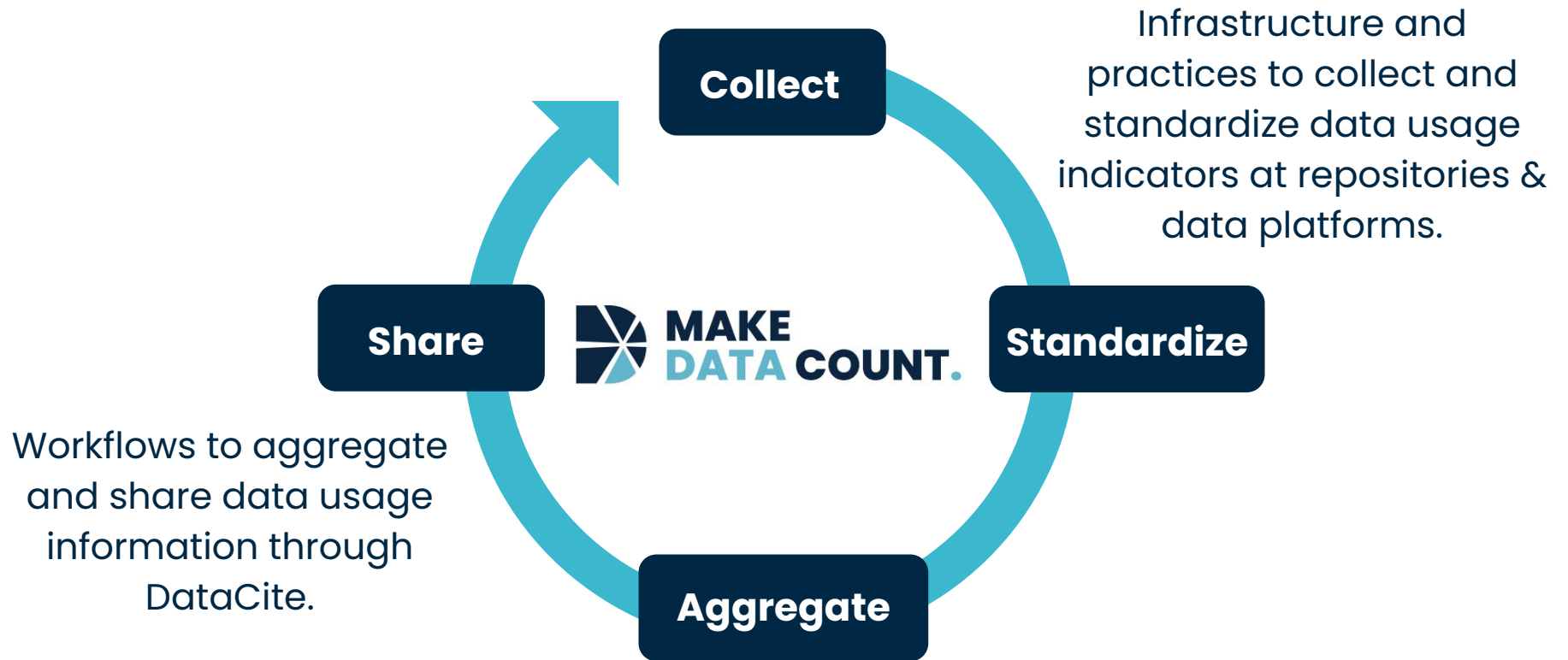


Standards & evidence
to contextualize data
metrics



Advocacy to promote
recognition of data
usage & impact in
assessment

Collecting, standardizing, and sharing data usage



Collecting, standardizing & sharing data usage counts

01 Implement MDC recommendations

Two approaches to collect data views & downloads normalized per the COUNTER standard:

- Usage tracker widget that collects counts on dataset landing pages.
- Log processing server wide.

02 Share the data usage counts

- Display usage counts on dataset records to raise visibility & recognition for data creators.
- Submission of counts to DataCite enables aggregation & discoverability.

03 Review & gain insights

- Review usage counts for alignment with usage tracking standards or potential outliers.
- Gain insights into highly-used datasets and repository use over time.

Make Data Count infrastructure allows the community to collect and access standardized counts of data views and downloads.

The image shows two overlapping screenshots of dataset landing pages. The top screenshot is from DataCite Commons, displaying the dataset 'Health condition data for Platypus from New South Wales and Victoria' with 186 views and 16 downloads. The bottom screenshot is from DRYAD, showing the same dataset with 172 views and 16 downloads. Both pages list the creators: Jana Stewart, Gilad Bino, Tahneal Hawke, and Richard Kingsford, all from UNSW Sydney.

DataCite Commons Page:

- Dataset: Health condition data for Platypus from New South Wales and Victoria
- DOI: <https://doi.org/10.5061/dryad.br15dv9d>
- Usage: 186 Views, 16 Downloads
- Creators: Jana Stewart (UNSW Sydney), Gilad Bino (UNSW Sydney), Tahneal Hawke (UNSW Sydney), Richard Kingsford
- Cite as: Stewart, J., Bino, G., Hawke, T., & Kingsford, R. (2021). Health condition data for Platypus from New South Wales and Victoria (Version 6) [Data set]. Dryad. <https://doi.org/10.5061/DRYAD.BRV15DV9D>

DRYAD Page:

- Dataset: Health condition data for Platypus from New South Wales and Victoria
- DOI: <https://doi.org/10.5061/dryad.br15dv9d>
- Usage: 172 views, 16 downloads, 1 citations
- Author affiliations: Stewart, Jana; Bino, Gilad; Hawke, Tahneal; Kingsford, Richard
- Published: Nov 29, 2021 on Dryad
- Data files: Nov 29, 2021 version files (51.99 KB), Platypus_Health_Data_NSW VIC_2015-2018.xlsx (51.99 KB)
- Subject keywords: Biological sciences
- Funding: Australian Research Council (LP150100093)

commons.datacite.org/doi.org/10.5061/dryad.br15dv9d

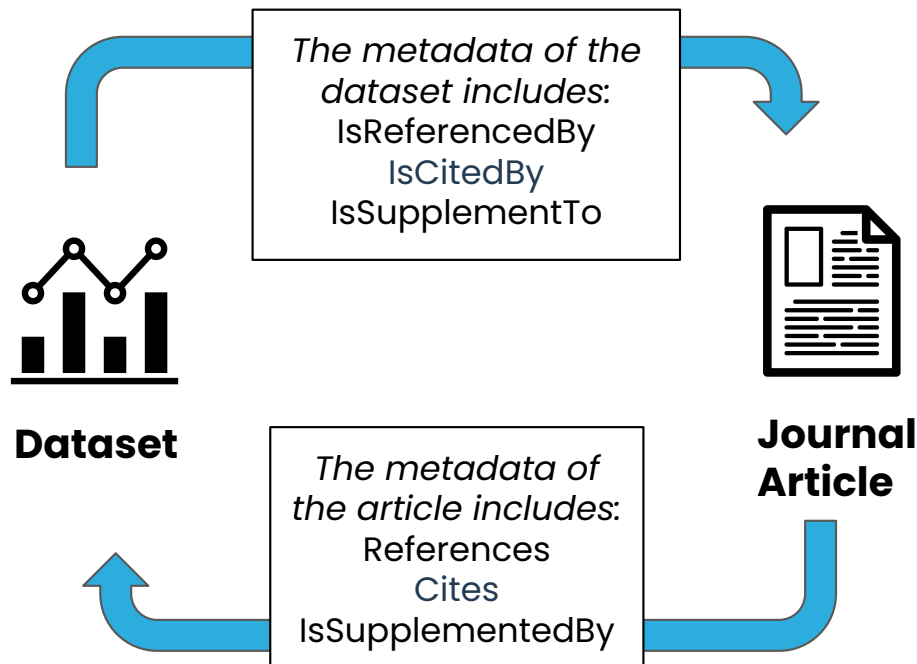
Collecting, standardizing & sharing data citations

Make Data Count infrastructure enables collection and access to data citations.

The **RelatedIdentifier property** in the DataCite metadata schema creates links between objects.

Sub-properties:

- relatedIdentifierType
- relationType



Collecting, standardizing & sharing data citations

Make Data Count infrastructure enables collection and access to data citations.



NAHDAP
National Addition & HIV Data Archive Program

Int/Create Account Find Data Deposit Data Train

National Survey on Drug Use and Health, 2009 (ICPSR 29621)

Version Date: Nov 23, 2015 [Cite this study](#) | [Share this page](#)

Principal Investigator(s):
United States Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies


Series:

- National Survey on Drug Use and Health (NSDUH) Series

<https://doi.org/10.3886/ICPSR29621.v6>

```
"relatedIdentifiers": [  
  {  
    "relationType": "IsCitedBy",  
    "relatedIdentifier": "10.2105/AJPH.2017.303743",  
    "resourceTypeGeneral": "JournalArticle",  
    "relatedIdentifierType": "DOI"  
  }  
],
```

DataCite JSON



Type to search... [Pages](#) [Support](#) [Sign In](#)

[Works](#) [People](#) [Organizations](#) [Repositories](#)

National Survey on Drug Use and Health, 2009 <https://doi.org/10.3886/icpsr29621>

[Download Metadata](#) **10 Citations** [Add to ORCID Record](#)

[Sign in to add this work to your ORCID record.](#)

Cite as

United States Department of Health And Human Services. Substance Abuse And Mental Health Services Administration. Office Of Applied Studies. (2010). *National Survey on Drug Use and Health, 2009* (Version v0) [Data set]. ICPSR - Interuniversity Consortium for Political and Social Research. <https://doi.org/10.3886/ICPSR29621>

Description Other Identifiers Creators Funders Registration

The National Survey on Drug Use and Health (NSDUH) series (formerly titled National Household Survey on Drug Abuse) primarily measures the prevalence and correlates of drug use in the United States. The surveys are designed to provide quarterly, as well as annual, estimates. Information is provided on the use of illicit drugs, alcohol, and tobacco among members of United States households aged 12 and older. Questions included age at first use as well as lifetime, annual, and past-month usage for the following drug classes: marijuana, cocaine (and crack), hallucinogens, heroin, inhalants, alcohol, tobacco, and nonmedical use of prescription drugs, including pain relievers, tranquilizers, stimulants, and sedatives. The survey covered substance abuse treatment history and perceived need for treatment, and included questions from the Diagnostic and Statistical Manual (DSM) of Mental Disorders that allow diagnostic criteria to be applied. The survey included questions concerning treatment for both substance abuse and mental health-related disorders. Respondents were also asked about personal and family income sources and amounts, health care access and coverage, illegal activities and arrest record, problems resulting from the use of drugs, and needle-sharing. Questions introduced in previous administrations were retained in the 2009 survey, including questions asked only of respondents aged 12 to 17. These "youth experiences" items covered a variety of topics, such as neighborhood environment, illegal activities, drug use by friends, social support, extracurricular activities, exposure to substance abuse prevention and education programs, and perceived adult attitudes toward drug use and activities such as school work. Several measures focused on prevention-related themes in this section. Also retained were questions on mental health and access to care, perceived risk of using drugs, perceived availability of drugs, driving and personal behavior, and cigar smoking. Questions on the tobacco brand used most often were introduced with the 1999 survey. For the 2008 survey, Adult mental health questions were added to measure symptoms of psychological distress in the worst period of distress that a person experienced in the past 30 days and

commons.datacite.org/doi.org/10.3886/icpsr29621

We lack a full picture of data citations

Existing processes do not yet provide the full picture on the use of data.



Only a fraction of data citations are reported as structured references.



Workflows needed to capture citations for different data identifiers.



Citations identified by different groups are stored in different (and sometimes closed) locations.

The Data Citation Corpus



A large open resource of data citations.

The Corpus brings together citations identified by different methodologies, including persistent identifier metadata, curation, and full-text mining.

- ✓ Multiple sources: DataCite Event Data, Chan Zuckerberg Initiative, ASAP, EuropePMC
- ✓ Datasets with DOIs & accession numbers
- ✓ Transparency on provenance

**The Data Citation Corpus
aggregates 10 million data citations**

The screenshot shows the Zenodo interface for the 'Data Citation Corpus Data File'. The page includes a search bar, navigation links for 'Communities' and 'My dashboard', and a 'Log in' / 'Sign up' button. The dataset is titled 'Data Citation Corpus Data File' and is associated with 'DataCite' and 'Make Data Count'. It was published on August 15, 2025, as version v4.1. The page displays statistics: 7K views and 1K downloads. A table compares 'All versions' and 'This version' for views, downloads, and data volume. The 'This version' has 670 views, 193 downloads, and a data volume of 205.5 GB. The page also lists the sources of the data citations: DataCite Event Data, Chan Zuckerberg Initiative (CZI) Science Knowledge Graph, Aligning Science Across Parkinson's (ASAP), and Europe PMC.

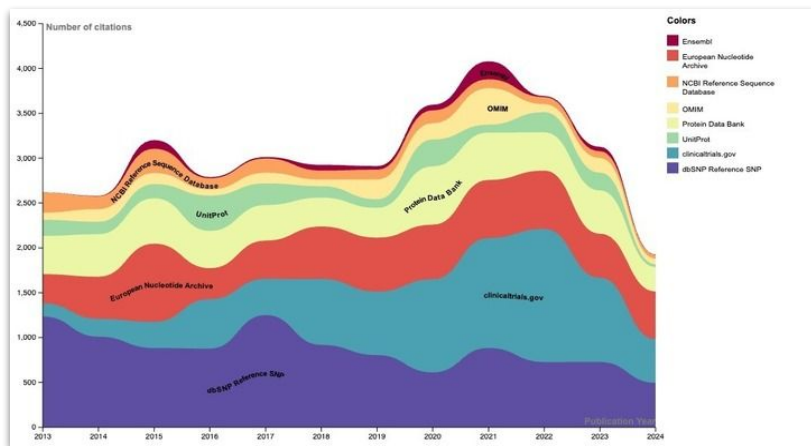
	All versions	This version
Views	7,202	670
Downloads	1,264	193
Data volume	1.3 TB	205.5 GB

Corpus data file:
<https://doi.org/10.5281/zenodo.11196858>



Data usage insights for libraries

Analysis of data citations for Northwestern University & the University of Colorado, Boulder, using data citations in the Data Citation Corpus and in Europe PMC.



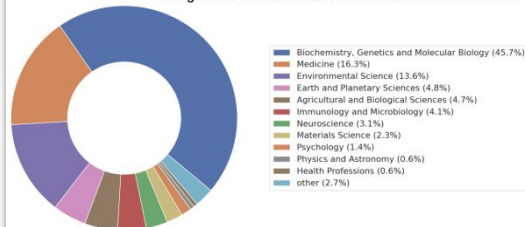
Most used repositories: dbSNP, Protein Data Bank, European Nucleotide Archive

Data-intensive research areas at Northwestern University: cancer, immunology, infectious diseases, biochemistry, molecular biology, neuroscience.

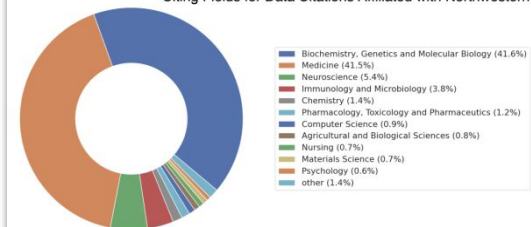
Data-intensive research areas at University of Colorado, Boulder:

environmental sciences (microbial ecology, polar research), molecular biology, genetics, plant sciences.

Citing Fields for Data Citations Affiliated with CU Boulder



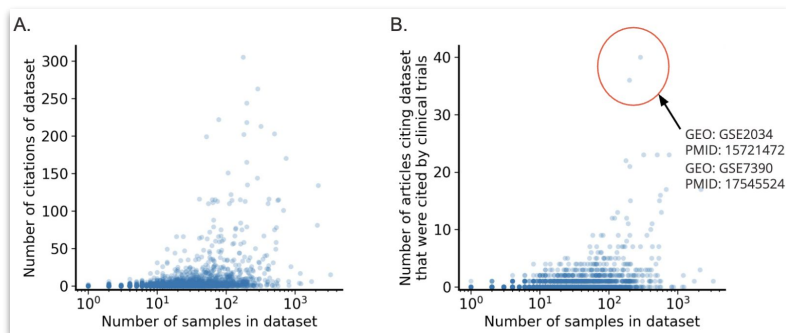
Citing Fields for Data Citations Affiliated with Northwestern



Wittenberg, J., Portenoy, J., Puebla, I., & Holmes, K. (2025). Automating data citation at scale to advance open data metrics. Association of College & Research Libraries (ACRL 2025), Minneapolis. Zenodo. <https://doi.org/10.5281/zenodo.15130354>

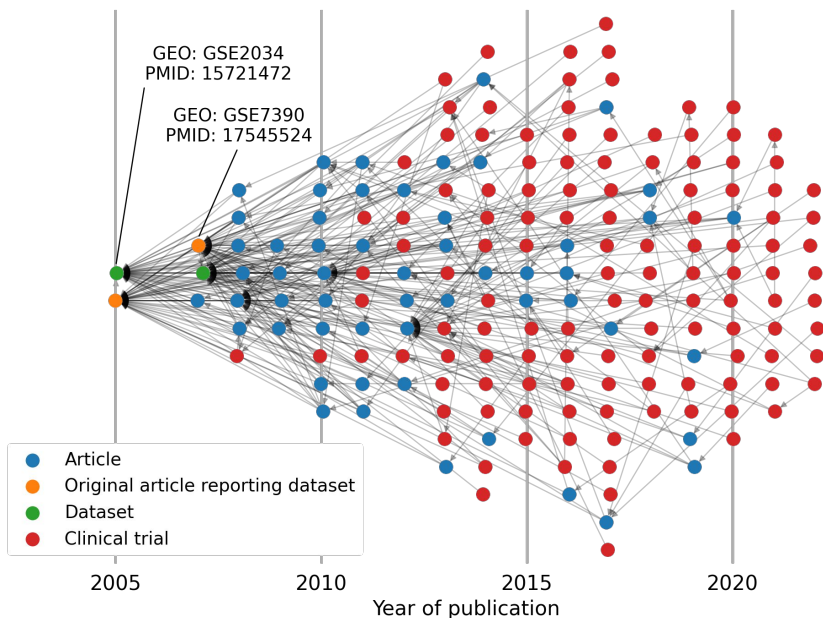
Insights into translational data impact

Collaboration with researchers at Northwestern University: The Data Citation Corpus as a basis to trace how datasets move from basic science into health applications.



Identified two GEO datasets for human data with a high number of citations.

Richardson, R., Puebla, I., Portenoy, J., Gutzman, K., & Holmes, K. (2025). Opening up translational data impact through the Data Citation Corpus. Zenodo. <https://doi.org/10.5281/zenodo.15299655>



Citation networks between the GEO datasets and the articles reporting the datasets, and the publications that subsequently cited those articles. 55 non-clinical trial articles cite GSE2034 & GSE7390, these were later cited by 124 clinical trials.

Advancing data evaluation at institutions

A partnership between HELIOS Open and Make Data Count, the **'Implementing data evaluation in academia' Working Group** has developed resources to support implementation of data evaluation in institutional processes.

Implementation guide

- Sample language for tenure & promotion policies
- Researcher CV template
- Guidance for review committees

01

Tenure & Promotion Policies

Sample Language

This section contains sample text that may be included in policy documents to signal recognition of data contributions and open scholarship practices. Institutions are encouraged to adapt and extend the examples as appropriate to their needs. The article 'Ten simple rules for recognizing data and software contributions in hiring, promotion, and tenure' provides further guidance and considerations for recognizing data and software in institutional processes.

Research productivity and open scholarship

Our institution recognizes research practices designed to make research more reliable, rigorous, and aligned to open scholarship. The evaluation process is transparent and values the broad spectrum of research products that researchers produce and openly share as part of their research. When describing their research productivity, faculty are encouraged to document and share open datasets, software workflows, protocols, code, and other scholarly products that contribute to research transparency, accessibility, and impact, alongside traditional research outputs such as journal publications.

Assessments of research productivity at our institution incorporate attributes related to open data and open research outputs, such as those below:

- Creation or curation of datasets that are shared openly to the extent that this is ethically and legally possible.
- Creation and open sharing of research or analysis tools, workflows, software or code.
- Creation of resources for facilitating research, such as data collections, or platforms for hosting and/or interacting with large datasets to facilitate collaborative research. This also includes platforms to interact with closed data that cannot be shared in their raw form due to privacy or ethical concerns.

Open scholarship practices

Faculty who adopt open scholarship practices – including data management and sharing, open access publications and preprints, open methods and workflows (e.g., preregistration), open educational materials, and open code – are recognized for their contributions to scientific rigor and reproducibility.

Assessment of research contributions

The evaluation of research contributions includes consideration toward evidence of adherence to community standards for open and reproducible scholarship, complete reporting of activities across the research cycle (e.g., via preregistration and adherence to reporting guidelines) and the development and open sharing (where ethically and legally permitted) of research tools, instruments, code, and data.

Evaluators consider the time, complexity, and collaborative nature of producing high-quality reusable datasets and data infrastructures. This includes curating, maintaining, and sharing datasets ethically and sustainably. Evaluators may also want to recognize contributions that enhance accessibility in research, particularly work that addresses the specific needs of underrepresented communities.

Research that involves the use of existing datasets or collections of datasets is also considered, and researchers are encouraged to discuss these approaches and how they relate to practices in their field in their personal statements for consideration by the evaluators.

Evaluation of potential for impact

The evaluation of the potential for impact of the research contributions can be informed by a wide range of indicators that span beyond traditional publication metrics. Indicators of potential for impact may include quantitative measures, contextual information, and qualitative evidence. Such indicators may include:

- Complete reporting of results and open sharing of research outputs associated with published articles, through preregistrations, open data, code and/or materials
- Counts of open datasets and other open outputs
- Citations to data, software or preregistrations
- Dataset views and downloads
- Software downloads and installs
- Use of the data, software or open outputs in projects, activities, or collaborations that advanced understanding in the field or addressed knowledge gaps
- Evidence of adherence to community standards for open scholarship (e.g. FAIR principles) and standards for conducting ethically sound and reproducible research
- Evidence of community engagement or endorsement of the open data or other open outputs developed

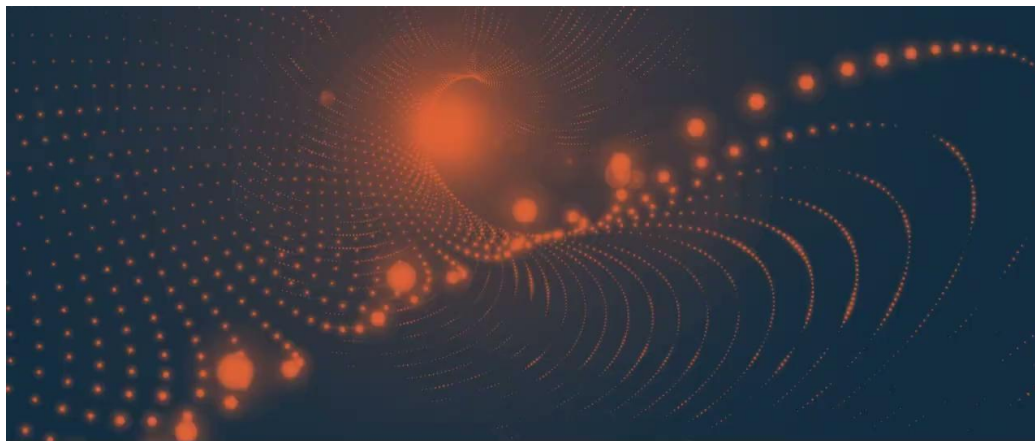
The outline of the potential for impact for the research contributions can take a narrative format where the researcher provides a contextualized description of their key data and open outputs and how they advanced the generation of new ideas, tools, methodologies or knowledge. This outline can include a description of the pathways to impact that the researcher envisions for their data and open outputs, including possible future uses that might advance research endeavours, spur innovation, or bring societal benefit.

Advancing data evaluation at institutions



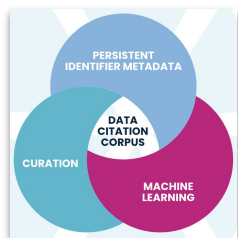
A partnership between HELIOS Open and Make Data Count, the **'Implementing data evaluation in academia' Working Group** has developed resources to support implementation of data evaluation in institutional processes.

- Institutional case studies
- Maturity model for data evaluation to be released this month



<https://makedatacount.org/explore-resources/institutions/>

Engage with Make Data Count



Explore the **Data Citation Corpus** data file and share your feedback.

Are you collecting data citations? Contribute these to the Corpus.



Get in touch about ways to **implement usage tracking** at your repository.



Share your data evaluation examples and perspectives, we'll be happy to amplify your story.

**Thank you
Danke**



**MAKE
DATA COUNT.**

info@makedatacount.org

makedatacount.org

Zenodo community: zenodo.org/communities/makedatacount