GREI

Generalist Repository Ecosystem Initiative

GREI Data citation best practices for repositories

Introduction

One of the objectives of the Generalist Repository Ecosystem Initiative (GREI) is to implement data metrics that enable reporting on the reach and impact of NIH-funded research data.

Data citations are a key component of the measures of data usage, as they bring benefits to the data creators, the data users, and the scholarly communication ecosystem more broadly:

- Data citations are a signal of a dataset being used in research (beyond mere exploration), providing valuable information to evaluate data usage.
- Data citations provide credit for the data producer, the citation recognizes the individual(s) or organization(s) that collected and shared the data used in the citing work.
- For academic researchers, accruing citations to datasets can also be valuable as part of research evaluation frameworks (e.g. for hiring or promotion), as they provide evidence of the reach of their open datasets.
- Surveys of researchers regularly show that getting citations to their research papers as well as citations to the datasets themselves are among the biggest motivators for them to publish their data (see for example The State of Open Data report).
- Data citations increase the rigor and reproducibility of research, enabling data users to document the source of the data they employed as part of research activities.
- Data citations can enable the development of tools to aid search and discovery of research works; increasing visibility on what outputs cite a dataset can help researchers find other scholarly objects relevant to their work.

Providing visibility for data citations is therefore a way of increasing the information on data usage available to the community, it signals the added value of data repositories, and can create an incentive for researchers to share more of their data and to cite open data.

The importance of data citations has been recognized by earlier community work, for example, in the form of the FORCE11 Data Citation Principles, which cover the purpose, function and attributes of citations to data. The principles note as part of its guidance that while citations may vary in style, citations to data "*should be included in the full reference list along with citations to other types of works.*"

All of the GREI repositories (Dataverse, Dryad, Figshare, Open Science Framework, Mendeley Data, Vivli, and Zenodo) already collect data citations or have it on their roadmap to add this feature. This group of repositories has also been considering approaches to capture those citations in a manner that provides transparent and consistent information to the community on citations for data.

We outline here, from our perspective as generalist repositories, our recommendations for handling data citations in repositories. We hope that this will be useful for other repositories that are seeking to implement or update their practices for data citations.

What do we mean by data citation?

Citations can be created for any scholarly output, but for the scope of this document, we are focusing specifically on citations to datasets. For the purpose of our recommendations, we refer to a data citation as:

A link between a dataset in a repository and another scholarly output where both have persistent identifiers.

Examples of data citations include those below (please note this is not an exhaustive list):

- A journal article that cites the dataset underlying the results
- A preprint that cites a dataset re-analysed using a different methodology
- A software tool that cites the training dataset used during development
- A conference presentation or poster that cites the dataset produced during the study
- A clinical guideline cites a dataset that informed the recommendations
- A derived dataset (e.g. harmonized data, an analysis-ready dataset) that cites the original dataset on which it builds

Collecting data citations

Repositories can collect citations to datasets through self report by authors as they deposit and/or update data records, or harvest data citations from external sources. Data sources employed by GREI repositories to harvest data citations are listed below:

- DataCite / Crossref

- Dimensions
- Europe PMC
- NASA ADS (Astrophysics Data System)

We recommend that repositories submit data citations they collect to DataCite (see DataCite's documentation for contributing data citations to DataCite) so that these connections can be aggregated and discoverable by others in the community.

Storing data citations

Repositories collect data citations via the metadata for the datasets they host. We recommend that repositories use the metadata fields recommended by GREI to collect data citations, specifically, those listed below to establish the relationship between the dataset and the citing object:

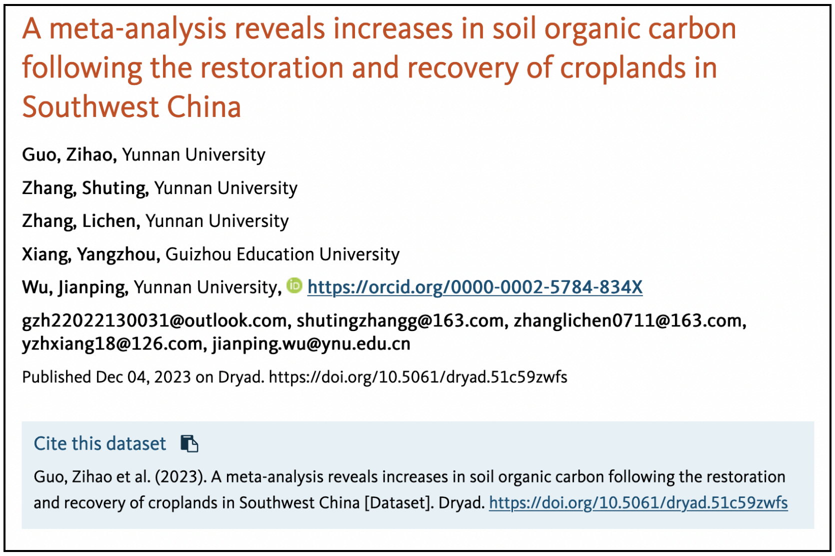
12. relatedIdentifier
12.a relatedIdentifierType
12.b relationType
12.f resourceTypeGeneral

The above fields align to the properties in the DataCite metadata schema for establishing citations. Note that under DataCite's schema, the following relationType entities as part of the metadata for the dataset are recorded as a data citation:

IsCitedBy
IsReferencedBy
IsSupplementTo

'Cite As' template for data citation

In order to encourage researchers and other parties to cite datasets they use, data repositories should provide a citation template on the landing page of the dataset. We include below a citation template example from the Dryad repository:



A meta-analysis reveals increases in soil organic carbon following the restoration and recovery of croplands in Southwest China

Guo, Zihao, Yunnan University

Zhang, Shuting, Yunnan University

Zhang, Lichen, Yunnan University

Xiang, Yangzhou, Guizhou Education University

Wu, Jianping, Yunnan University, https://orcid.org/0000-0002-5784-834X

gzh22022130031@outlook.com, shutingzhangg@163.com, zhanglichen0711@163.com, yzhxiang18@126.com, jianping.wu@ynu.edu.cn

Published Dec 04, 2023 on Dryad. https://doi.org/10.5061/dryad.51c59zwfs

Cite this dataset

Guo, Zihao et al. (2023). A meta-analysis reveals increases in soil organic carbon following the restoration and recovery of croplands in Southwest China [Dataset]. Dryad. https://doi.org/10.5061/dryad.51c59zwfs

The NLM Style Guide recommends that citations of datasets include the following elements: Author, Title, Type of Medium, Publisher, Date of Publication Date, Date of Update/Revision, Date of Citation[1], Availability (i.e. url or DOI). An example citation is included below[2]:

Di Stefano B, Collombet S, Graf T. Time-resolved gene expression profiling during reprogramming of C/EBPα-pulsed B cells into iPS cells [dataset]. 2014 May 22 [cited 2014 Oct 6]. In: figshare [Internet]. Available from: http://dx.doi.org/10.6084/m9.figshare.939408

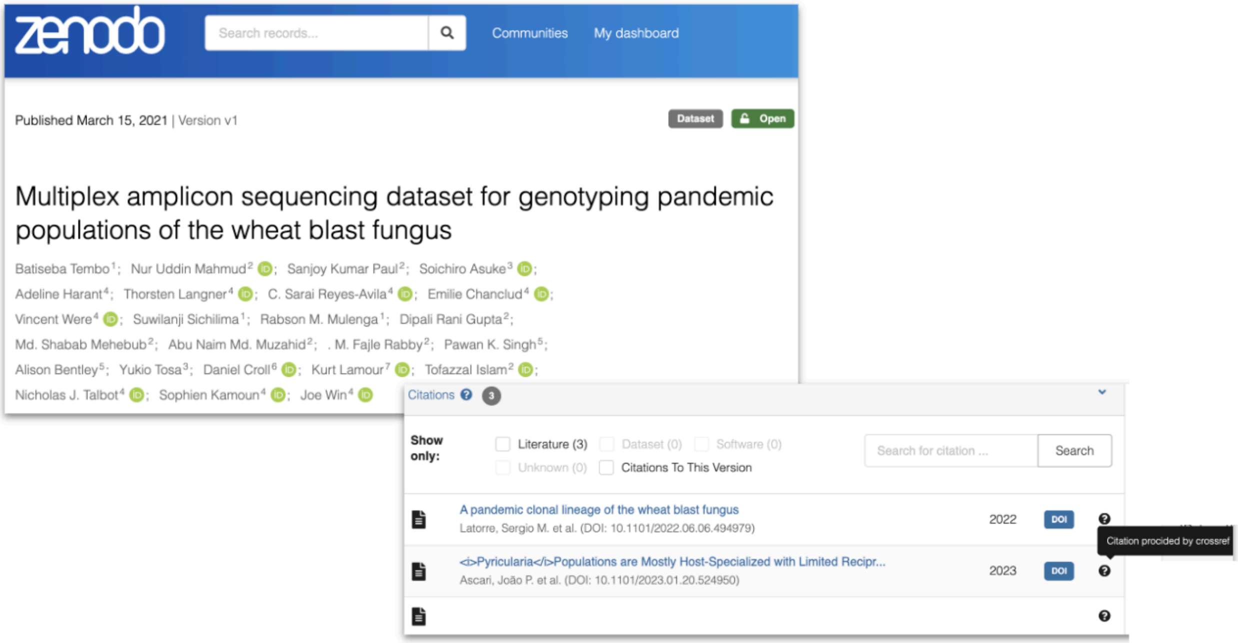
Exposing data citations

Repositories should expose the citations for individual datasets on the landing page for the dataset record. To increase transparency and trust in the information displayed for data citations, we recommend that repositories indicate the provenance for the data citation, i.e. the source for the asserted citation. The citation information listed at the repository should include:

- The number of citations to the dataset
- The list of citing objects and their associated identifiers, or a link to where such a list of citing objects can be accessed
- A breakdown of the citations according to their provenance, in text format - where the repository harvests data citations from external sources (see examples in section Collecting data citations), those should be listed

[1] Note the date of citation will only be created at the time at which the citation is introduced into the reference list, and thus, it will not be part of the template provided by the repository.

[2] Citation adapted from https://www.ncbi.nlm.nih.gov/books/NBK7273/#A57904.



Published March 15, 2021 | Version v1

Multiplex amplicon sequencing dataset for genotyping pandemic populations of the wheat blast fungus

Batiseba Tembo[1]; Nur Uddin Mahmud[2]; Sanjoy Kumar Paul[2]; Soichiro Asuke[3]; Adeline Harant[4]; Thorsten Langner[4]; C. Sarai Reyes-Avila[4]; Emilie Chanclud[4]; Vincent Were[4]; Suwilanji Sichilima[1]; Rabson M. Mulenga[1]; Dipali Rani Gupta[2]; Md. Shabab Mehebub[2]; Abu Naim Md. Muzahid[2]; . M. Fajle Rabby[2]; Pawan K. Singh[5]; Alison Bentley[5]; Yukio Tosa[3]; Daniel Croll[6]; Kurt Lamour[7]; Tofazzal Islam[2]; Nicholas J. Talbot[4]; Sophien Kamoun[4]; Joe Win[4]

Citations 3

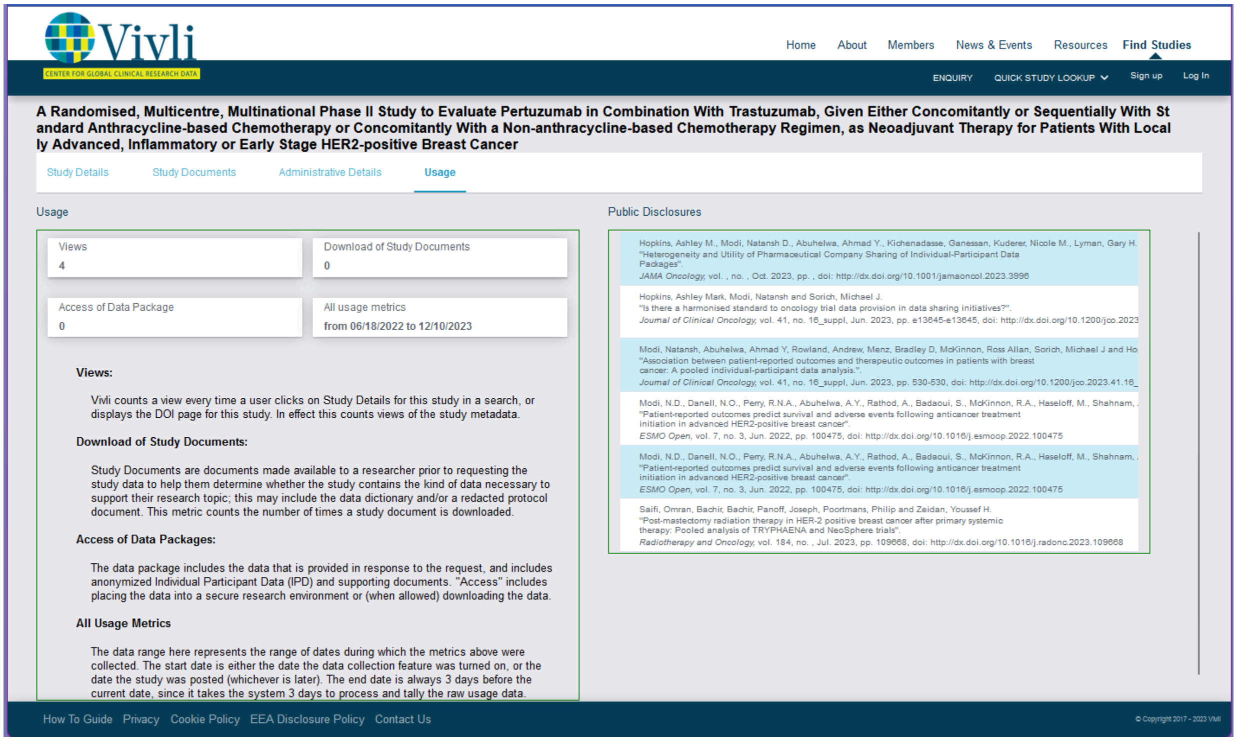
Show only: □ Literature (3) □ Dataset (0) □ Software (0) □ Unknown (0) □ Citations To This Version

A pandemic clonal lineage of the wheat blast fungus
Latorre, Sergio M. et al. (DOI: 10.1101/2022.06.06.494979) 2022 DOI

<i>Pyricularia</i>Populations are Mostly Host-Specialized with Limited Recipr...
Ascari, João P. et al. (DOI: 10.1101/2023.01.20.524950) 2023 DOI

Citation procided by crossref

Data citations for dataset on Zenodo: https://doi.org/10.5281/zenodo.4605959.



A Randomised, Multicentre, Multinational Phase II Study to Evaluate Pertuzumab in Combination With Trastuzumab, Given Either Concomitantly or Sequentially With St andard Anthracycline-based Chemotherapy or Concomitantly With a Non-anthracycline-based Chemotherapy Regimen, as Neoadjuvant Therapy for Patients With Local ly Advanced, Inflammatory or Early Stage HER2-positive Breast Cancer

Usage

Views 4	Download of Study Documents 0
Access of Data Package 0	All usage metrics from 06/18/2022 to 12/10/2023

Views:

Vivli counts a view every time a user clicks on Study Details for this study in a search, or displays the DOI page for this study. In effect this counts views of the study metadata.

Download of Study Documents:

Study Documents are documents made available to a researcher prior to requesting the study data to help them determine whether the study contains the kind of data necessary to support their research topic; this may include the data dictionary and/or a redacted protocol document. This metric counts the number of times a study document is downloaded.

Access of Data Packages:

The data package includes the data that is provided in response to the request, and includes anonymized Individual Participant Data (IPD) and supporting documents. "Access" includes placing the data into a secure research environment or (when allowed) downloading the data.

All Usage Metrics

The data range here represents the range of dates during which the metrics above were collected. The start date is either the date the data collection feature was turned on, or the date the study was posted (whichever is later). The end date is always 3 days before the current date, since it takes the system 3 days to process and tally the raw usage data.

Public Disclosures

Hopkins, Ashley M., Modi, Natansh D., Abuhelwa, Ahmad Y., Kichenadasse, Ganessan, Kuderer, Nicole M., Lyman, Gary H.
"Heterogeneity and Utility of Pharmaceutical Company Sharing of Individual-Participant Data Packages".
JAMA Oncology, vol. , no. , Oct. 2023, pp. , doi: http://dx.doi.org/10.1001/jamaoncol.2023.3996

Hopkins, Ashley Mark, Modi, Natansh and Sorich, Michael J.
"Is there a harmonised standard to oncology trial data provision in data sharing initiatives?".
Journal of Clinical Oncology, vol. 41, no. 16_suppl, Jun. 2023, pp. e13645-e13645, doi: http://dx.doi.org/10.1200/jco.2023

Modi, Natansh, Abuhelwa, Ahmad Y, Rowland, Andrew, Menz, Bradley D, McKinnon, Ross Allan, Sorich, Michael J and Ho
"Association between patient-reported outcomes and therapeutic outcomes in patients with breast cancer: A pooled individual-participant data analysis.".
Journal of Clinical Oncology, vol. 41, no. 16_suppl, Jun. 2023, pp. 530-530, doi: http://dx.doi.org/10.1200/jco.2023.41.16_

Modi, N.D., Danell, N.O., Perry, R.N.A., Abuhelwa, A.Y., Rathod, A., Badaoui, S., McKinnon, R.A., Haseloff, M., Shahnam,
"Patient-reported outcomes predict survival and adverse events following anticancer treatment initiation in advanced HER2-positive breast cancer".
ESMO Open, vol. 7, no. 3, Jun. 2022, pp. 100475, doi: http://dx.doi.org/10.1016/j.esmoop.2022.100475

Modi, N.D., Danell, N.O., Perry, R.N.A., Abuhelwa, A.Y., Rathod, A., Badaoui, S., McKinnon, R.A., Haseloff, M., Shahnam,
"Patient-reported outcomes predict survival and adverse events following anticancer treatment initiation in advanced HER2-positive breast cancer".
ESMO Open, vol. 7, no. 3, Jun. 2022, pp. 100475, doi: http://dx.doi.org/10.1016/j.esmoop.2022.100475

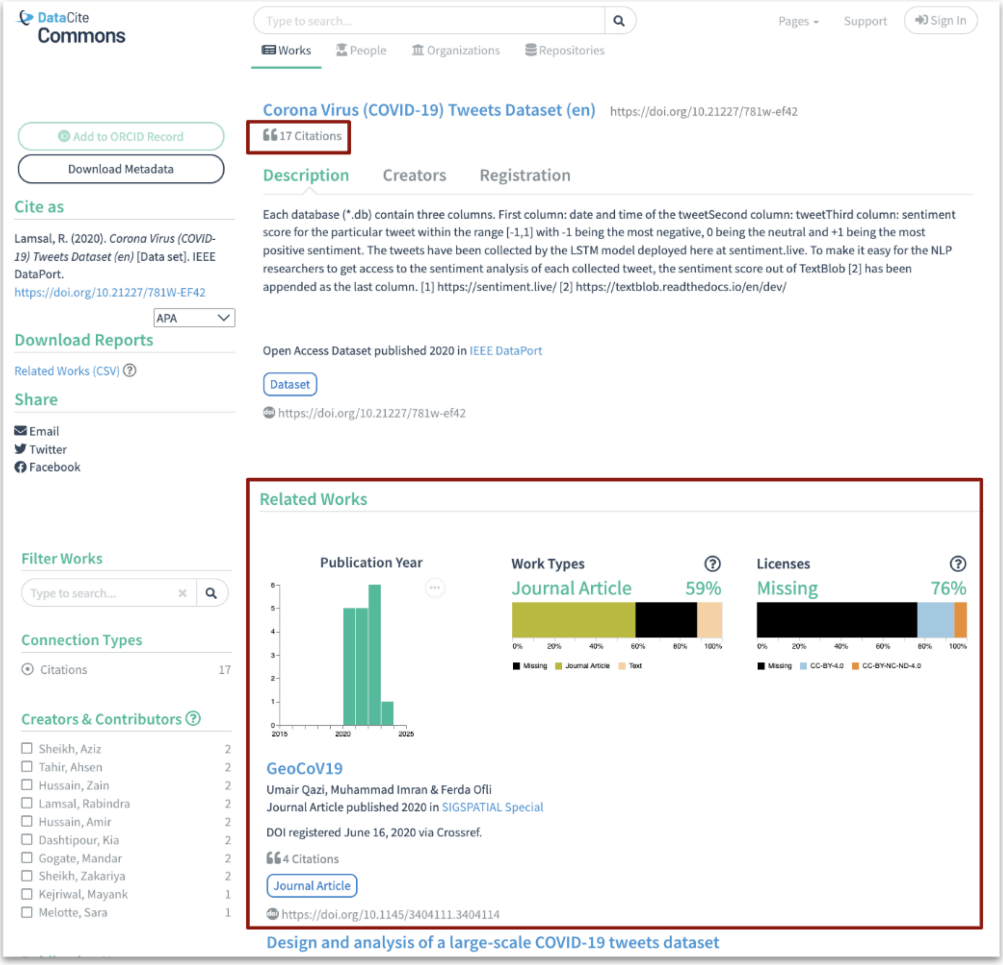
Saifi, Omran, Bachir, Bachir, Panoff, Joseph, Poortmans, Philip and Zeidan, Youssef H.
"Post-mastectomy radiation therapy in HER-2 positive breast cancer after primary systemic therapy: Pooled analysis of TRYPHAENA and NeoSphere trials".
Radiotherapy and Oncology, vol. 184, no. , Jul. 2023, pp. 109668, doi: http://dx.doi.org/10.1016/j.radonc.2023.109668

Data citations for dataset on Vivli, displayed in the 'Public Disclosures' box for the dataset at https://doi.org/10.25934/00005758. The term "Public Disclosures" was chosen at a time when the term "Citations" had a narrower connotation than the definition above.

The repository should also document, as part of the information pages for contributors and users, information on the mechanisms by which the repository collects and stores data citation, including the sources it employs.

Data citations on DataCite Commons

Data citations captured via the metadata deposit with DataCite, as well as some citations available from Crossref metadata, are also displayed on the DataCite portal DataCite Commons. This web interface enables searches for scholarly resources with persistent identifiers as well as connections to metadata provided by DataCite, Crossref, ORCID, and others. DataCite Commons provides the number of citations for the dataset as well as a list of citing objects.



Corona Virus (COVID-19) Tweets Dataset (en) https://doi.org/10.21227/781w-ef42

17 Citations

Description · Creators · Registration

Each database (*.db) contain three columns. First column: date and time of the tweetSecond column: tweetThird column: sentiment score for the particular tweet within the range [-1,1] with -1 being the most negative, 0 being the neutral and +1 being the most positive sentiment. The tweets have been collected by the LSTM model deployed here at sentiment.live. To make it easy for the NLP researchers to get access to the sentiment analysis of each collected tweet, the sentiment score out of TextBlob [2] has been appended as the last column. [1] https://sentiment.live/ [2] https://textblob.readthedocs.io/en/dev/

Open Access Dataset published 2020 in IEEE DataPort

Dataset

https://doi.org/10.21227/781w-ef42

Cite as

Lamsal, R. (2020). *Corona Virus (COVID-19) Tweets Dataset (en)* [Data set]. IEEE DataPort.
https://doi.org/10.21227/781W-EF42

Download Reports

Related Works (CSV)

Share

Filter Works

Connection Types

○ Citations	17

Creators & Contributors

□ Sheikh, Aziz	2
□ Tahir, Ahsen	2
□ Hussain, Zain	2
□ Lamsal, Rabindra	2
□ Hussain, Amir	2
□ Dashtipour, Kia	2
□ Gogate, Mandar	2
□ Sheikh, Zakariya	2
□ Kejriwal, Mayank	1
□ Melotte, Sara	1

Related Works

GeoCoV19
Umair Qazi, Muhammad Imran & Ferda Ofli
Journal Article published 2020 in SIGSPATIAL Special

DOI registered June 16, 2020 via Crossref.

4 Citations

Journal Article

https://doi.org/10.1145/3404111.3404114

Design and analysis of a large-scale COVID-19 tweets dataset

DataCite Commons record for https://doi.org/10.21227/781w-ef42, displaying the number of citations and citing objects.

Data Citation Corpus

There have been challenges in consistently collecting information about data citations across repositories and the literature, due to the fact that workflows (e.g. by publishers) have not been optimized to collect citations to datasets, and that different parties are using their own mechanisms to capture citations and store that information in different locations. This has limited the ability of stakeholders such as institutions and funders to incorporate data as part of their evaluation frameworks.

To address these challenges and scale the data citations that are consistently made available to the community, DataCite is working on the development of the Data Citation Corpus, which will provide a centralized resource that compiles data citations from a variety of sources, and make data citation information readily and openly available to the community.

The data sources for the data citation corpus include:

- Persistent identifier authorities: Sources that collect citations as part of their DOI registration workflow, such as DataCite and Crossref.
- Third-party aggregators: Sources that aggregate or discover citations through various techniques, such as full-text mining and curation. For example, the Chan Zuckerberg Initiative (CZI) has contributed data citations identified through the CZI Knowledge Graph, which mines the text of publications via a machine-learning algorithm.

The Data Citations Corpus will include data citations in DataCite; this incorporates citations deposited by DataCite-member repositories, including the GREI repositories. We recommend that all data repositories contribute their data citations to DataCite so that those citations can be integrated into the Data Citations Corpus.

DataCite will engage closely with GREI repositories during the development process of the corpus and invites other repositories and interested parties to get in touch if they are interested in contributing data citations to the corpus or in providing feedback as potential users of this resource.

Appendix 1: Additional reading/Resources

- Curtin, L., Feri, L., Gautier, J., Gonzales, S., Gueguen, G., Scherer, D., Scherle, R., Stathis, K., Van Gulick, A., & Wood, J. (2023). GREI Metadata and Search Subcommittee Recommendations_V01_2023-06-29. Zenodo. https://doi.org/10.5281/zenodo.8101957
- 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
- Fenner, M., Crosas, M., Grethe, J.S. *et al.* A data citation roadmap for scholarly data repositories. *Sci Data* 6, 28 (2019). https://doi.org/10.1038/s41597-019-0031-8
- Data Citation Corpus: https://makedatacount.org/data-citation/
- JATS4R Data Citations recommendations, Version 2.0: https://doi.org/10.3789/niso-rp-36-2020
- Science, Digital; Hahnel, Mark; Smith, Graham; Schoenenberger, Henning; Scaplehorn, Niki; Day, Laura (2023). The State of Open Data 2023. Digital Science. Report. https://doi.org/10.6084/m9.figshare.24428194.v1

Appendix 2: Description of citation metadata for human readers and contributors

To support understanding of how repositories collect and store citations via metadata, and how this is aggregated in DataCite, we include below descriptions for specific instances in metadata that designate a data citation or where a dataset cites another scholarly object.

Metadata for citation *The examples below reflect the properties within the dataset DOI metadata that designate a citation*	**Description**
Dataset DOI: 10.5066/F70Z71H2 <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType> <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">https://doi.org/10.1021/acs.est.7b00812 </relatedIdentifier>	The dataset is cited by the article with DOI 10.1021/acs.est.7b00812
Dataset DOI: 10.5281/zenodo.5112001 <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType> <relatedIdentifier relatedIdentifierType="DOI" relationType="IsCitedBy">https://doi.org/10.1101/2021.01.04.425314</relatedIdentifier>	The dataset is cited by the preprint with DOI 10.1101/2021.01.04.425314
Dataset DOI: 10.5281/zenodo.8091721	The dataset cites the preprint with

<resourceType resourceTypeGeneral="Dataset">Dataset</resourceType> <relatedIdentifier relatedIdentifierType="DOI" relationType="Cites">https://doi.org/10.1101/2022.08.09.503355</relatedIdentifier>	DOI 10.1101/2022.08.09.503355
Dataset DOI: 10.5281/zenodo.10149985 <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType> <relatedIdentifier relatedIdentifierType="DOI" relationType="IsReferencedBy">https://doi.org/10.1101/2023.10.03.560679</relatedIdentifier>	The dataset is cited by the preprint with DOI 10.1101/2023.10.03.560679
Dataset DOI: 10.5281/zenodo.8399018 <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType> <relatedIdentifier relatedIdentifierType="PMID" relationType="IsSupplementedBy">10.5194/egusphere-2023-2295 </relatedIdentifier>	The dataset is supplemented by the article with DOI 10.5194/egusphere-2023-2295
Dataset DOI 10.6084/m9.figshare.24565030 <resourceType resourceTypeGeneral="Dataset">Dataset</resourceType> <relatedIdentifier relatedIdentifierType="DOI" relationType="IsSupplementTo">10.1159/000534801</relatedIdentifier>	The dataset supplements the article with DOI 10.1159/000534801