GREI Generalist Repository Ecosystem Initiative

GREI Collaborative Workshop

Discovering and reusing data in generalist repositories

January 25, 2023

Hi! We're your hosts for today

Julie Wood
Senior Director
Vivli

Kelly Stathis
Technical Community Manager
DataCite

Let's meet the facilitators

Eric Olson

Product Manager
Center for Open Science (OSF)

Sara Gonzales

Senior Data Librarian, Northwestern University, Zenodo

Blaine Butler

Scientist/ Product Manager
Center for Open Science (OSF)

David Scherer

Customer Consultant, Pure and Research Data Management
Elsevier

Julian Gautier

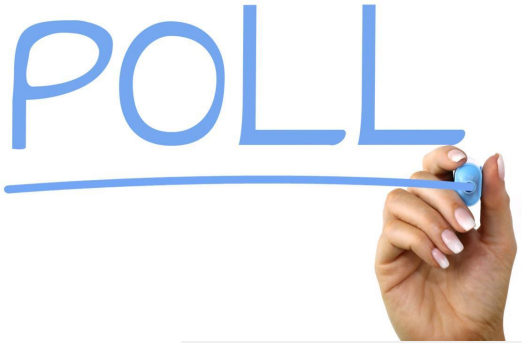
Product Research Specialist
Dataverse

Here's what we'll cover

- Welcome and introductions (5 min)
- What we heard yesterday in the training session (5 min)
- Why it is important to share data and its subsequent re-use (5 min)
- How persistent identifiers enable data discovery (25 min)
    - What are PIDs and DOIs?
    - How DOIs facilitate discovery
    - Metadata for data citation and reuse
- Data metrics that track re-use (5 minutes)

- Breakout sessions (30 min): facilitated session for questions, feedback, etc.

- Wrap-up and close (10 min)



Have you shared data or supported a researcher in sharing data?

Have you shared data or supported a researcher in sharing data?

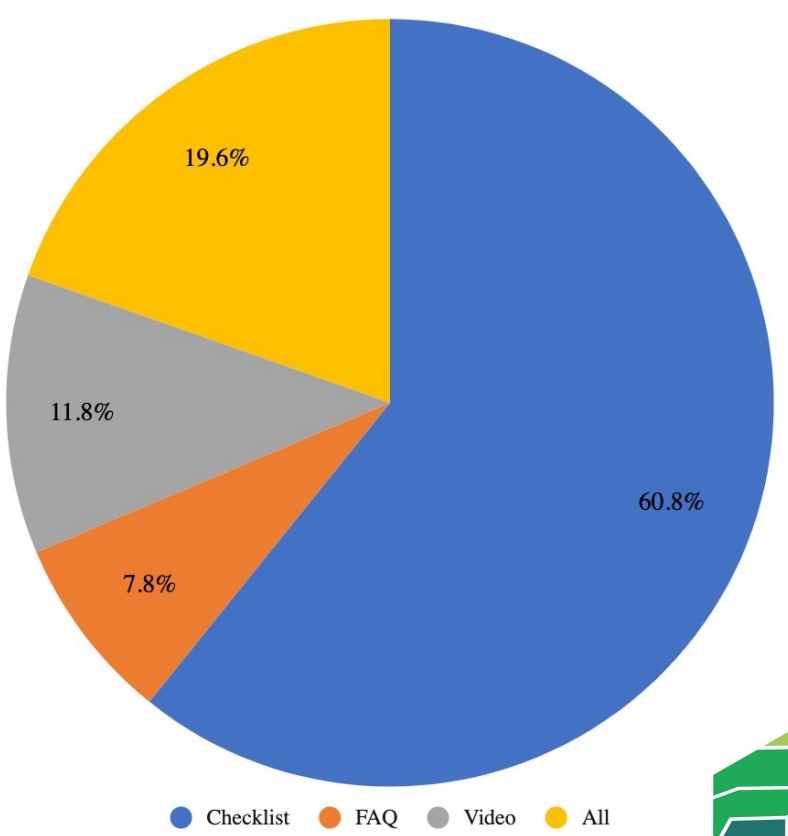
Yes, I have shared data — 14%

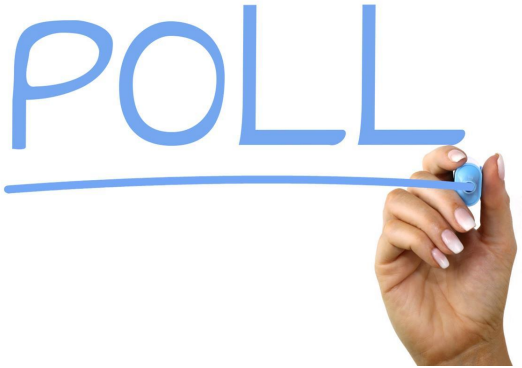
Yes, I have supported a researcher share data — 37%

Yes, I have both shared data and supported other researchers with sharing data — 43%

No, I have not shared data or supported a researcher with data sharing — 23%

Best practice is to review repository guidelines prior to initiating a submission. What is the most useful way for this information to be presented to make the process efficient? A submission checklist, instructional video, FAQs?





Based on what you've learned today, how do you feel about preparing data for a generalist repository?

Based on what you've learned today, how do you feel about preparing data for a generalist repository?

Completely prepared
3%

More prepared
97%

Less Prepared
0%

Why Is it important to share data

- Enables new discovery and new research questions through using and combining existing data with increased statistical power;
- Validates existing research results by peer review and reanalysis;
- Broadens research by enabling aggregation of data derived from disparate data generators
- Accelerates biomedical research discovery and innovations;
- Enhances research rigor and reproducibility
- Promote data reuse for future research studies

Ultimately sharing data speeds up the process of turning research results into knowledge, products and procedures to improve human health

Source: Susan Gregurick, NIH Dec. 2022
https://datascience.nih.gov/director/directors-blog-preparing-for-the-2023-data-management-and-sharing-policy

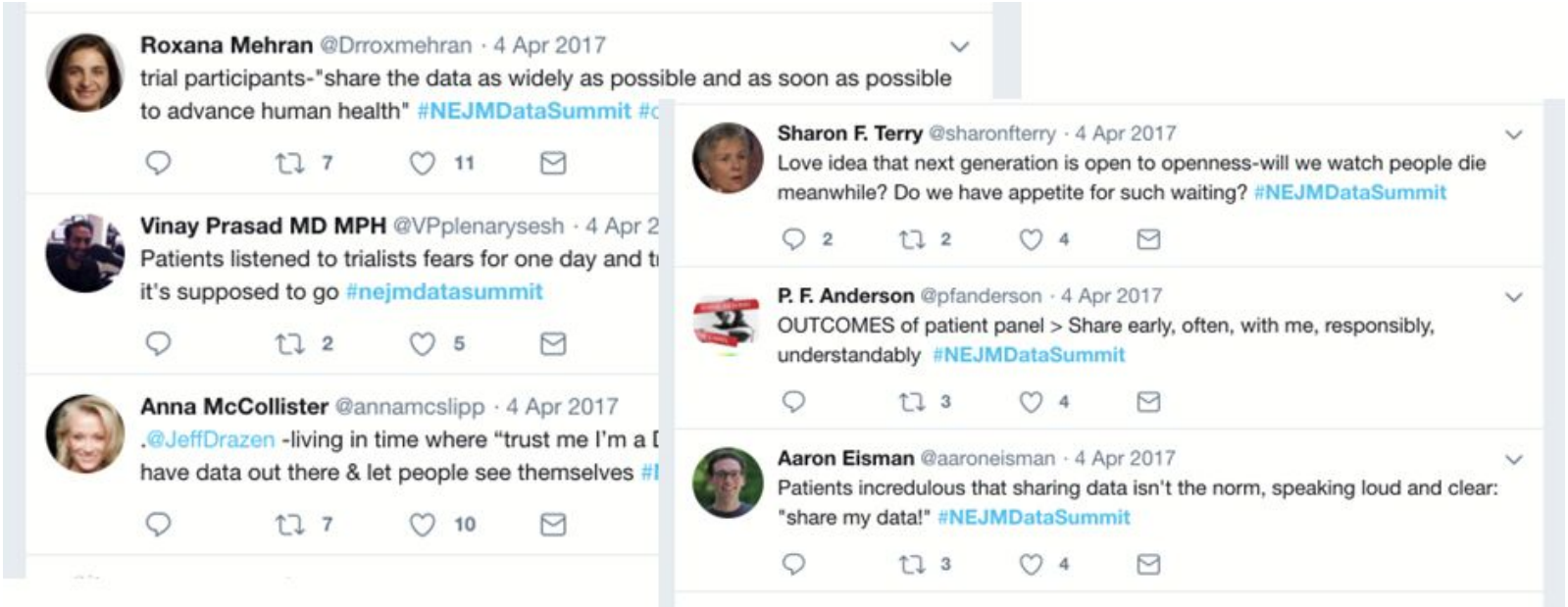
What Do Surveys Show Regarding Patient and Participant Preference Regarding Data Sharing?

- High levels of support for data sharing; however, patients are reluctant to have their data "commodified" purely for commercial gain
- If adequate safeguards were in place, trial participants are willing to share their data

Davidson S, McLean C, Treanor S, Aitken M, Cunningham-Burley S, Laurie G, et al. Public acceptability of data sharing between the public, private and third sectors for research purposes. Edinburgh: Scottish Government Social Research; 2013.Google Scholar

Mello, Michelle M., Van Lieou, and Steven N. Goodman. "Clinical trial participants' views of the risks and benefits of data sharing." *New England Journal of Medicine* 378.23 (2018): 2202-2211.

Participants expect data sharing and reuse



NEJM Aligning Incentives for Sharing Clinical Trial Data Summit, Boston, MA. April 2017

In 2-3 words, how do you currently search for datasets for secondary analysis (or support researchers in their search)?

Here's what we'll cover

- Welcome and introductions
- What we heard yesterday in the the training session (5 min)
- Why it is important to share data and its subsequent re-use (5 min)
- How persistent identifiers enable data discovery (20 min)
    - What are PIDs and DOIs?
    - How DOIs facilitate discovery
    - Metadata for data citation and reuse
    - Data metrics that track re-use

- Breakout sessions (25 min): facilitated session for questions, feedback, etc.

- Wrap-up (10 min)

DataCite

A persistent identifier (PID) is a unique, long-lasting reference to an entity.

Special URL that is registered in a known system, like DOI, ORCID or ROR

Always points to the same resource (or a metadata representation)

PID concepts

"**Identifier**" = a string of characters referring to an object

"**Unique**" = only refer to one object

"**Universal**" = are valid for the whole of the world (or world wide web)

"**Persistent**" = remain available

"**Actionable**" = a URL that resolves to a landing page with metadata information

"**Interoperable"** = integrate with other systems

"Open Metadata" = metadata availability through public and standardized APIs

"Strong Community" = sustainably funded and driven by the community they serve

PIDs for people, places, and things

PIDs for people (researchers) include ISNIs and ORCID iDs

https://orcid.org/0000-0001-6133-4045

PIDs for institutions (research organizations) including ROR

https://ror.org/03rmrcq20

PIDs for things (research outputs) include DOIs, handles, ARKs, and more

https://doi.org/10.17605/OSF.IO/KC4U9

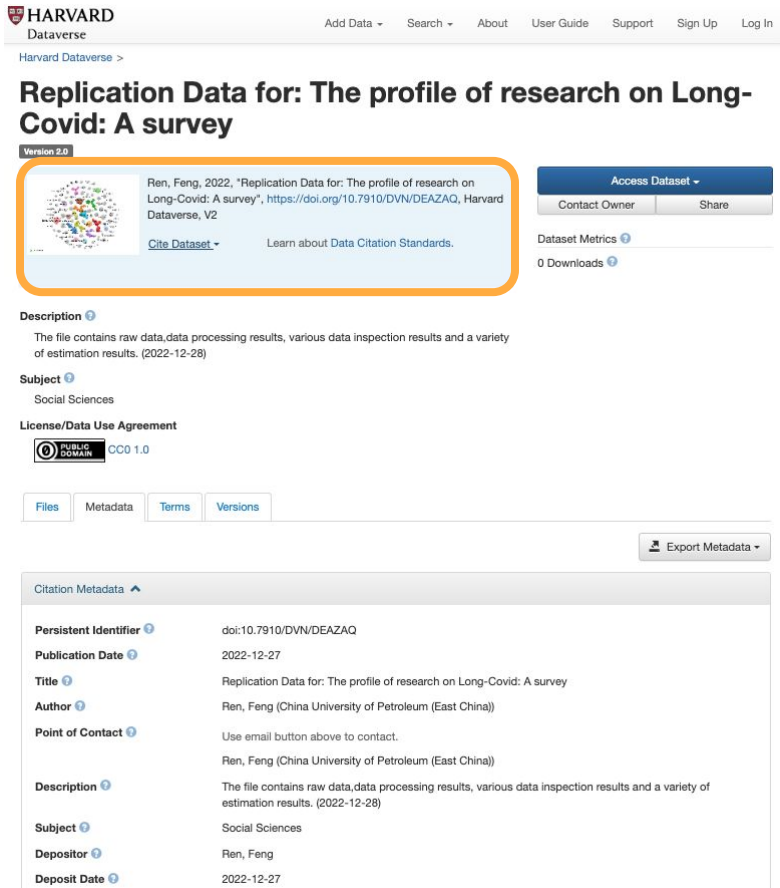
How do DOIs work?

- A DOI (Digital Object Identifier) uniquely identifies a resource
- Resolves to a URL—typically a landing page for the object
- Associated with accompanying standardized *metadata*
- Commonly used for research outputs
- DOI: prefix / suffix
- Formatted as a URL: https://doi.org/10.7910/DVN/DEAZAQ

Example: Dataset DOI

https://doi.org/10.7910/DVN/DEAZAQ

https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DEAZAQ



Example: Dataset DOI metadata

Required fields	**Example values**
Identifier	10.7910/DVN/DEAZAQ
Creators	Ren, Feng
Title	Replication Data for: The profile of research on Long-Covid: A survey
PublicationYear	2022
Publisher	Harvard Dataverse
ResourceType	Dataset

More DOI metadata

Some recommended/optional fields	**Description**
Subject	Subject, keyword, classification code, or key phrase describing the resource.
Description	E.g., an abstract for a dataset
Name identifier	Identifier for a creator (i.e., ORCID iD)
Affiliation and affiliation identifier	Creator's affiliation and associated identifier (i.e., ROR ID)
Rights	License (like CC-BY)
Funding Reference	Funder, award name and number

How are DOIs created?

DOIs for many types of research outputs—including data— are registered and maintained by repositories and publishers.

All GREI repositories register DataCite DOIs for datasets.

When a dataset is published:

- Dataset metadata is mapped to the DataCite Metadata Schema
- DOI is registered using this accompanying metadata
- DOI is displayed on the dataset landing page

How DOIs facilitate discovery:

DOI metadata is publicly available, searchable, and harvestable

Metadata is publicly available

DataCite DOI metadata is available to everyone in the public domain (CC0).

Where can you retrieve DataCite metadata?

- Through DataCite Commons
- Through search engines that use DataCite's APIs (e.g. REST API, GraphQL API) and OAI-PMH feed

DataCite Commons

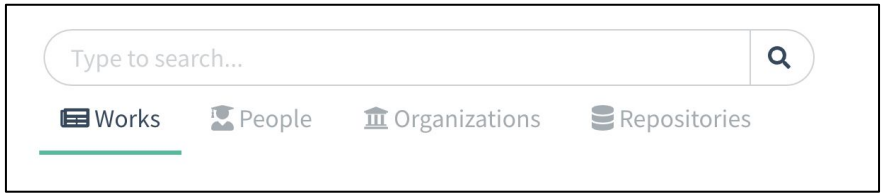
DataCite Commons (https://commons.datacite.org/) is a portal where anyone can go to search the entire DataCite metadata catalogue. You can find:

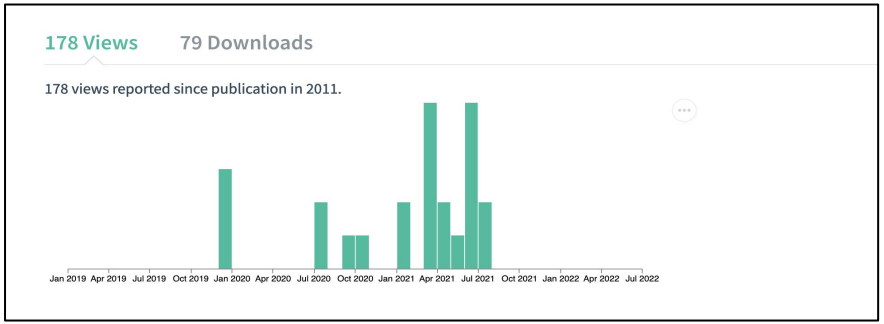
- Works (DOIs)
- Researchers
- Organizations
- Repositories
- Citations, views, downloads and more…

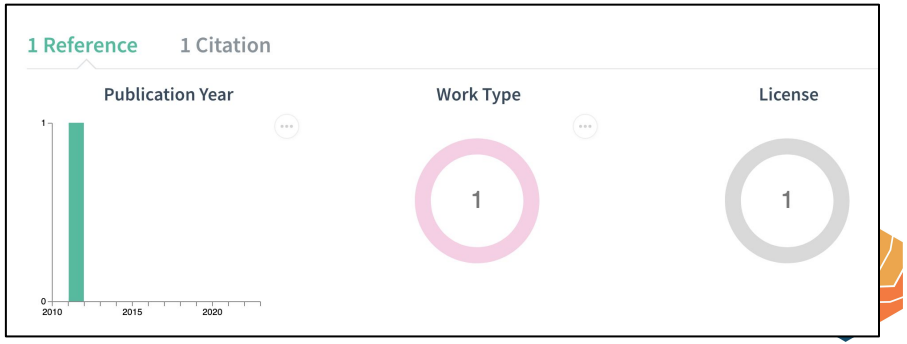
And they are all connected through links in the metadata.

DataCite Commons

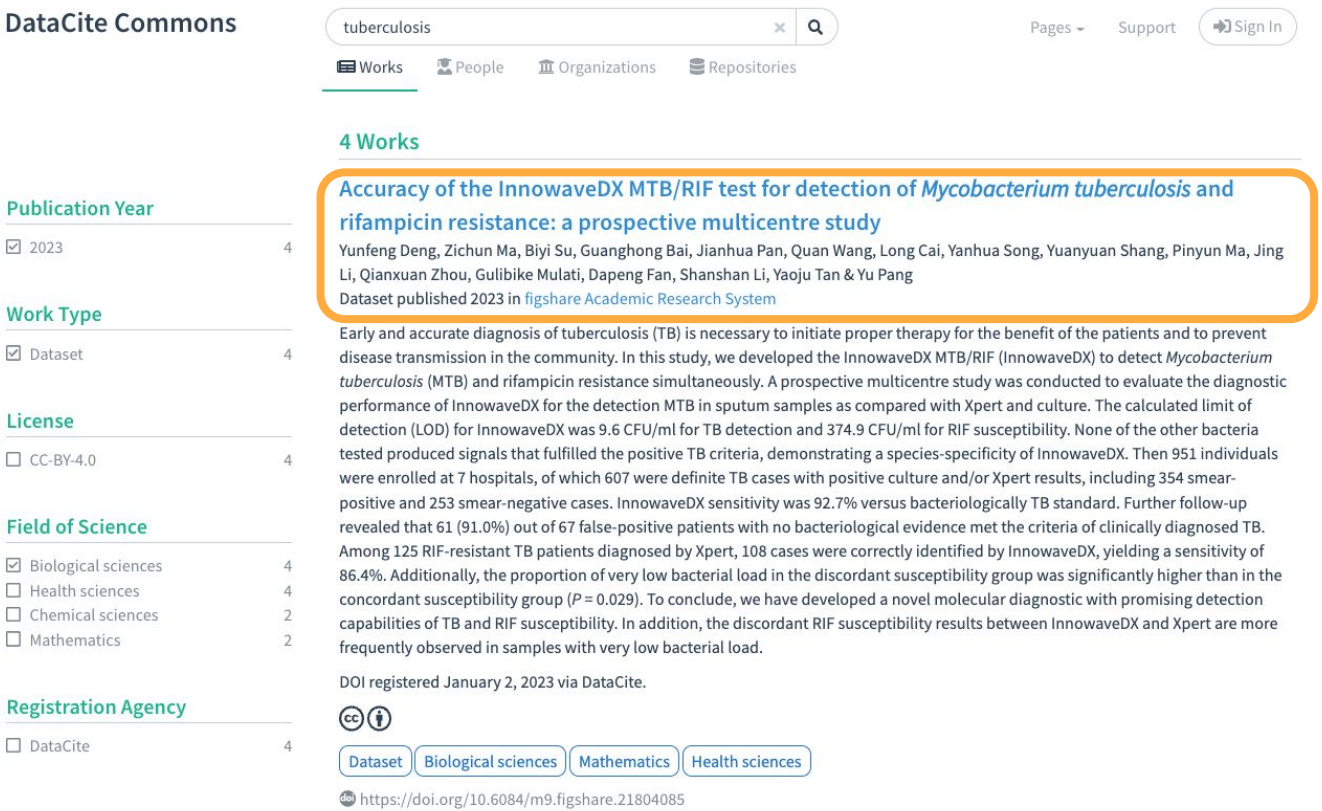
1. Search works, people, organizations and repositories
2. See related views and downloads
3. As well as citations and references





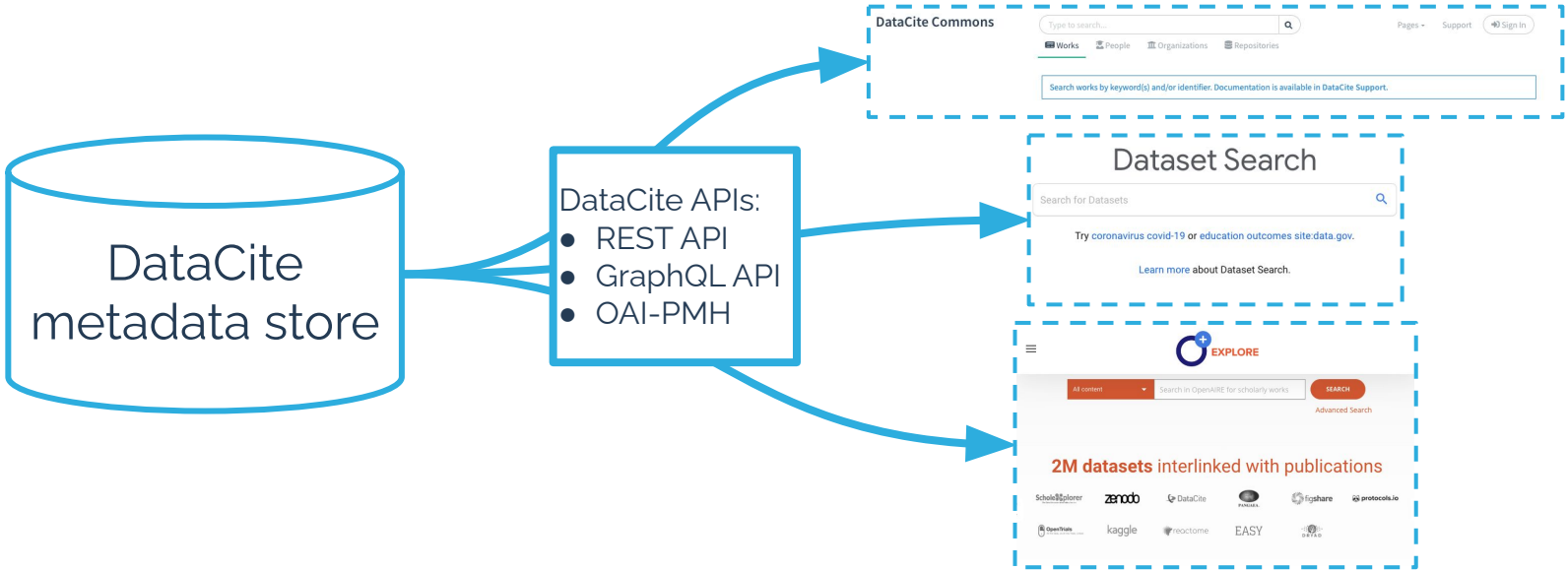


Example - search for datasets



Harvesters and aggregators

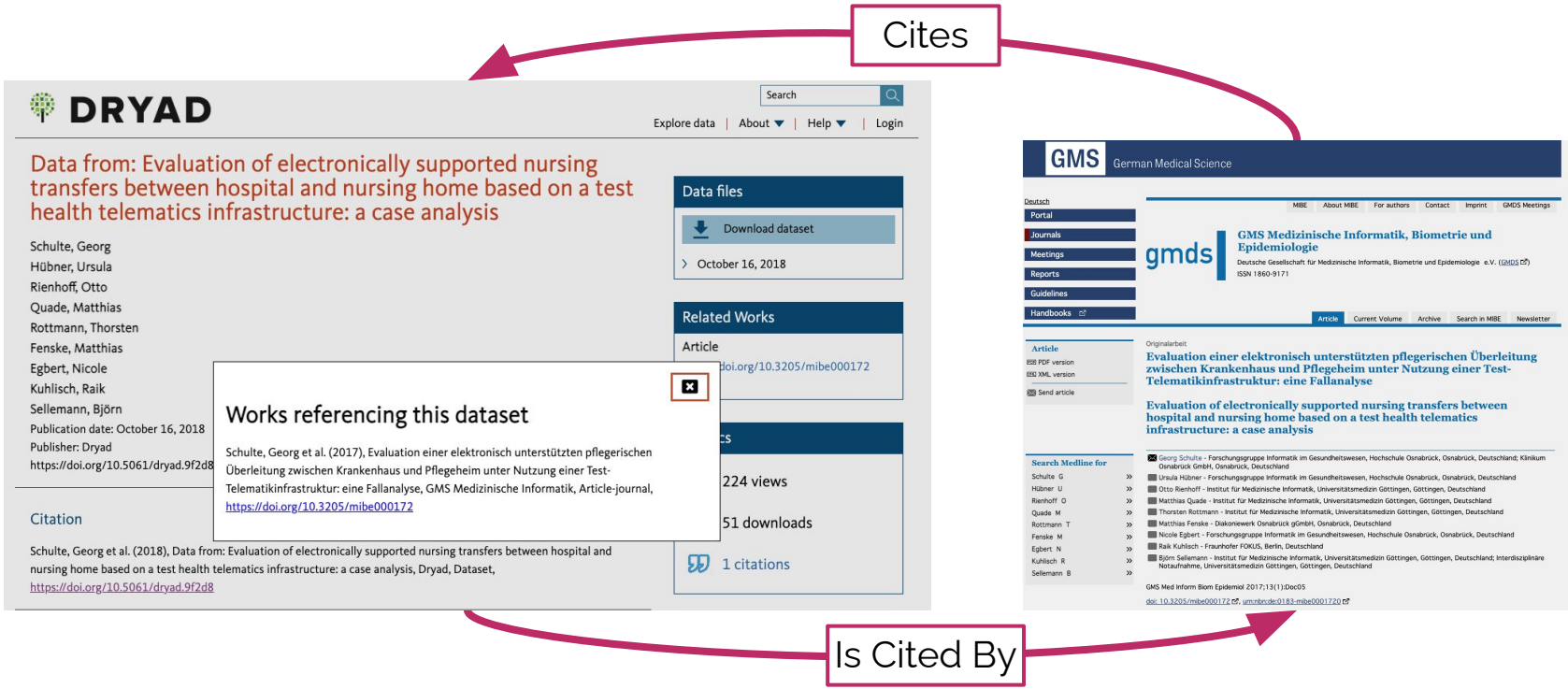
Metadata is harvested and made available by search engines that index metadata from DataCite, including OpenAIRE Explore, Clarivate's Data Citation Index (part of Web of Science) and more:



Why data citation?

John Borghi [@JohnBorghi]. (2021, May 5). As seen today at @CSVConference: Can I see the data related to your scientific paper? Another riff on that @XKCD comic. Original here: Https://xkcd.com/2456/ #csvconf #commallama [Tweet]. Twitter. https://twitter.com/JohnBorghi/status/1390033561579835393

Example: Data citation



Example: Data reuse

Cites

https://doi.org/10.5061/dryad.9f2d8

Data from: Evaluation of electronically supported nursing transfers between hospital and nursing home based on a test health telematics infrastructure: a case analysis

Georg Schulte, Ursula Hübner, Otto Rienhoff, Matthias Quade, Thorsten Rottmann, Matthias Fenske, Nicole Egbert, Raik Kuhlisch & Björn Sellemann

Version 1 of Dataset published 2018 in DRYAD

Background: Improper information transmission can lead to compromised patient safety and quality of life when patients are transferred from one setting to another. Electronic instruments may improve this situation, however, they are rarely used. Objective: The aim of this study therefore was to investigate the technical and organizational feasibility, usability, usefulness and completeness of an electronic instrument that is based on the German HL7 CDA standard for eNursing Summaries. Materials and methods: To this end, a test health telematics infrastructure, which included the German electronic health card, was established and nursing summary application was developed that allowed summary documents to be communicated between a hospital and a nursing home. The users were asked to evaluate the usability of the nursing summary application as well as to compare the usefulness and completeness of electronically and paper transmitted information. Results: This study demonstrated the feasibility of implementing an electronic nursing summary application that was based on the German HL7 CDA standard eNursing Summary and that was integrated in a test health telematics infrastructure. It could also be shown that the users rated this application as usable and that electronically supported patient transfers were superior to paper based ones. The use of the German electronic health card was regarded as a barrier by the users. Discussion: This study emphasizes the feasibility, relevance and barriers of electronically supported transfers of patients with nursing needs. Nurses working in hospitals and long-term care can integrate an application based on the HL7 CDA Standard ePflegebericht into their working processes and get better and more complete information. To ensure continuity of care in a sustainable manner in the future, the German HL7 CDA based eNursing Summary standard should become part of the German telematics infrastructure.

DOI registered October 16, 2017 via DataCite.

2 Citations 113 Views 23 Downloads

Dataset English

https://doi.org/10.5061/dryad.9f2d8

Challenges and care strategies associated with the admission to nursing homes in Germany: a scoping review

Stefanie Skudlik, Julian Hirt, Tobias D??ringer, Regina Thalhammer, Katharina L??ftl, Birgit Prodinger & Martin M??ller

Journal Article published 2023 in BMC Nursing

Other Identifiers

article-number: 5

DOI registered January 8, 2023 via Crossref.

Journal Article

https://doi.org/10.1186/s12912-022-01139-y

BMC Part of Springer Nature Search

BMC Nursing

Home About Articles Submission Guidelines Join The Board

Research | Open Access | Published: 05 January 2023

Challenges and care strategies associated with the admission to nursing homes in Germany: a scoping review

Stefanie Skudlik, Julian Hirt, Tobias Döringer, Regina Thalhammer, Katharina Lüftl, Birgit Prodinger & Martin Müller

BMC Nursing **22**, Article number: 5 (2023) | Cite this article

270 Accesses | Metrics

40. Schulte G, Hübner U, Rienhoff O, Quade M, Rottmann T, Fenske M, et al. Evaluation einer elektronisch unterstützten pflegerischen Überleitung zwischen Krankenhaus und Pflegeheim unter Nutzung einer Test-Telematikinfrastruktur: eine Fallanalyse. GMS Medizinische Informatik, Biometrie und Epidemiologie 2017. 10.5061/dryad.9f2d8

Have you referenced data? Did you use a persistent identifier to reference data? Will you in the future? (Or supported researchers in these practices)

Data usage and citation metrics

https://doi.org/10.5281/zenodo.3757476

COVID-19 CT Lung and Infection Segmentation Dataset

Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Minqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang & He Jian
Version Verson 1.0 of Content published 2020 in Zenodo

This dataset contains 20 labeled COVID-19 CT scans. Left lung, right lung, and infections are labeled by two radiologists and verified by an experienced radiologist.
To promote the studies of annotation-efficient deep learning methods, we set up three segmentation benchmark tasks based on this dataset https://gitee.com/junma11/COVID-19-CT-Seg-Benchmark. In particular, we focus on learning to segment left lung, right lung, and infections using pure but limited COVID-19 CT scans; existing labeled lung CT dataset from other non-COVID-19 lung diseases; heterogeneous datasets include both COVID-19 and non-COVID-19 CT scans.

DOI registered April 19, 2020 via DataCite.

28 Citations

Dataset

https://doi.org/10.5281/zenodo.3757476

Cite

APA

Jun, M., Cheng, G., Yixin, W., Xingle, A., Jiantao, G., Ziqi, Y., Minqing, Z., Xin, L., Xueyuan, D., Shucheng, C., Hao, W., Sen, M., Xiaoyu, Y., Ziwei, N., Chen, L., Lu, T., Yuntao, Z., Qiongjie, Z., Guoqiang, D., & Jian, H. (2020). *COVID-19 CT Lung and Infection Segmentation Dataset* (Verson 1.0) [Data set]. Zenodo. https://doi.org/10.5281/ZENODO.3757476

29 Citations

https://commons.datacite.org/doi.org/10.14291/tccon.ggg2014.ascension01.r0/1149285

Data Re-use Case study

Clinical trial data from from 4 data contributors lead to

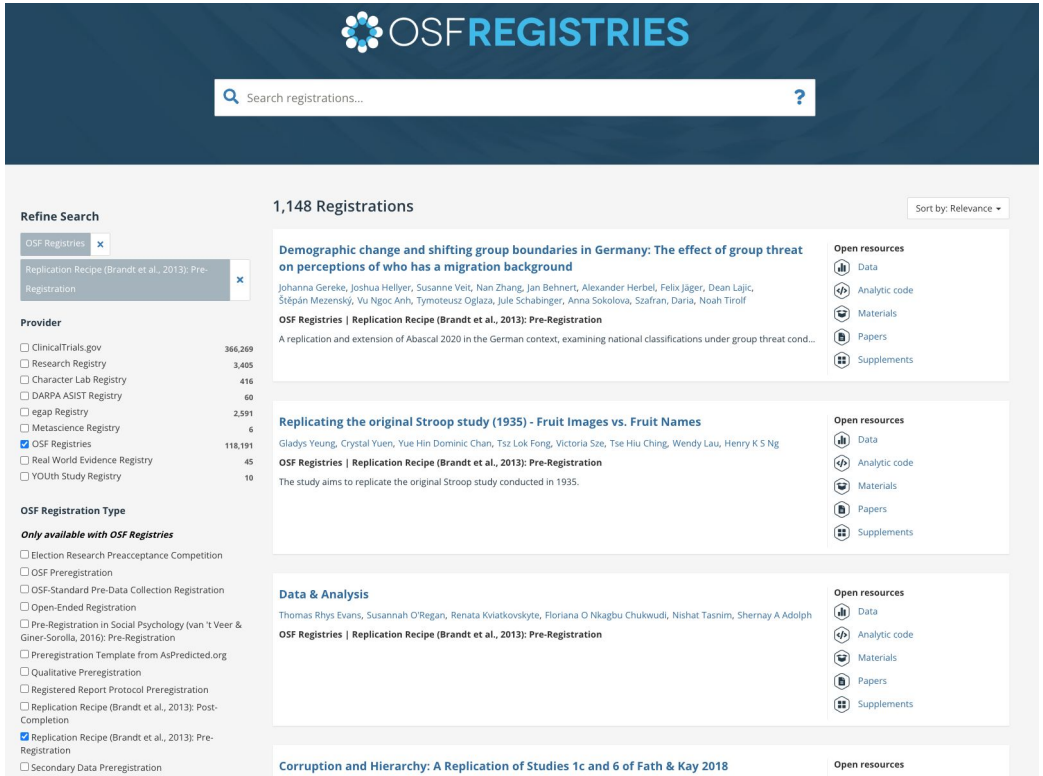
3 public disclosures

Khan FA, Stewart I, Saini G, Robinson KA, Jenkins RG. A systematic review of blood biomarkers with individual participant data meta-analysis of matrix-metalloproteinase-7 in IPF. Eur Respir J. 2021 Sep 29:2101612. doi: 10.1183/13993003.01612-2021 PMID: 34588192.

Reproducibility on the OSF

To facilitate convincing close replication attempts Brandt et al. developed a Replication Recipe, outlining standard criteria for a convincing close replication.

The OSF currently has **1,148** Registrations that have re-used data in order to perform a study to **Replicate Replication Recipe (Brandt et al., 2013)** with an additional **144** that have completed their study.

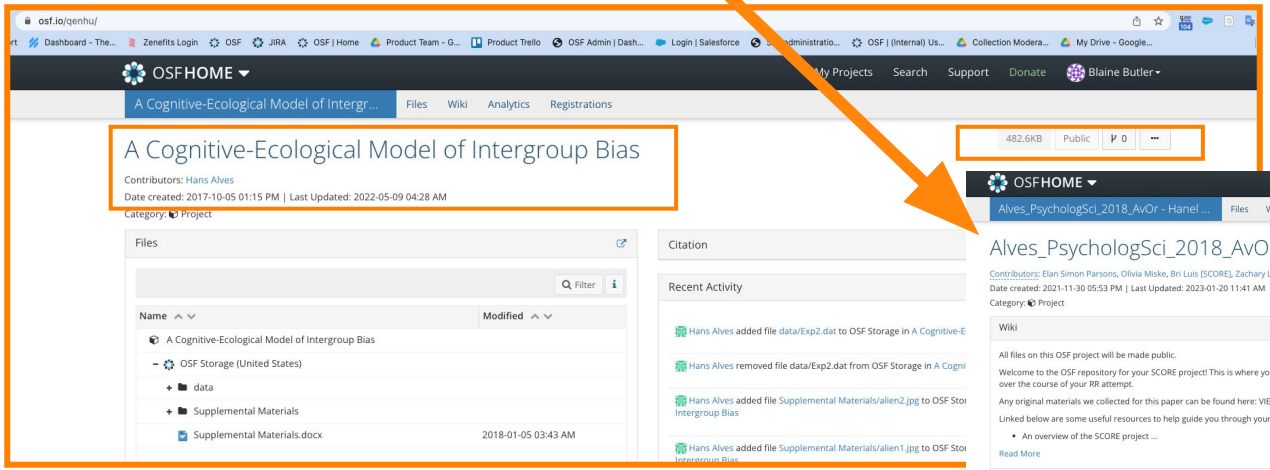


CENTER FOR OPEN SCIENCE

Re-use on the OSF

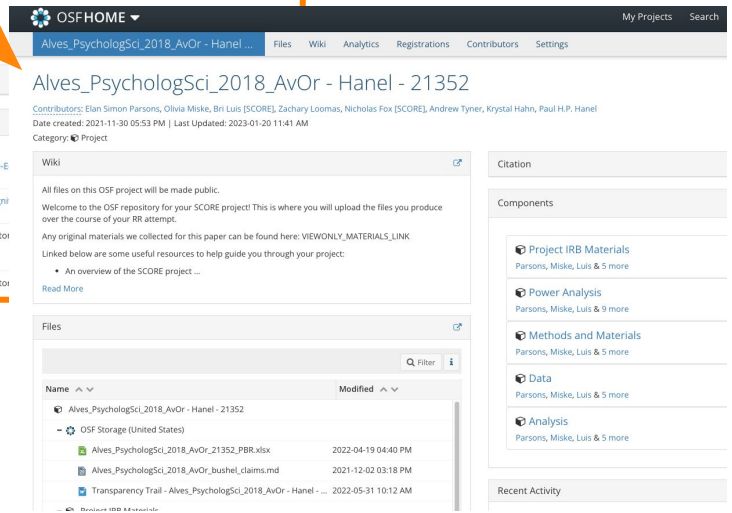
Systemizing Confidence in Open Research and Evidence(SCORE)

Initial author makes data publicly available
2017



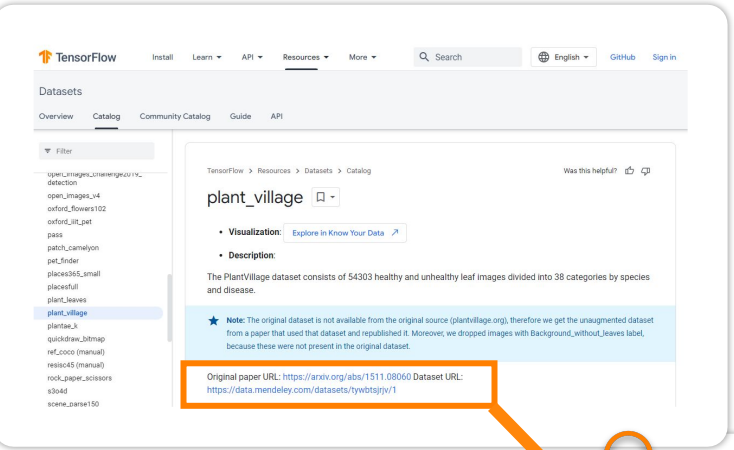
https://osf.io/qenhu

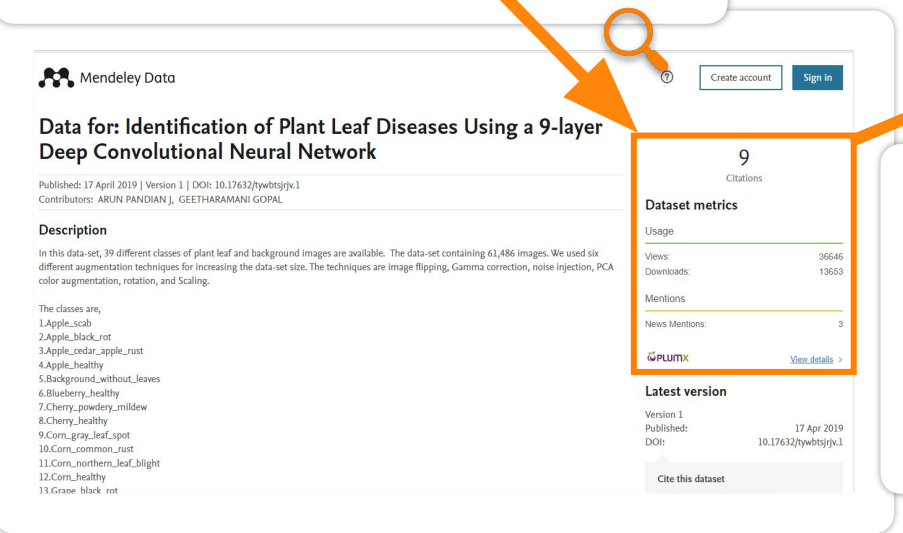
Data is then used to determine reproducibility claims
2021

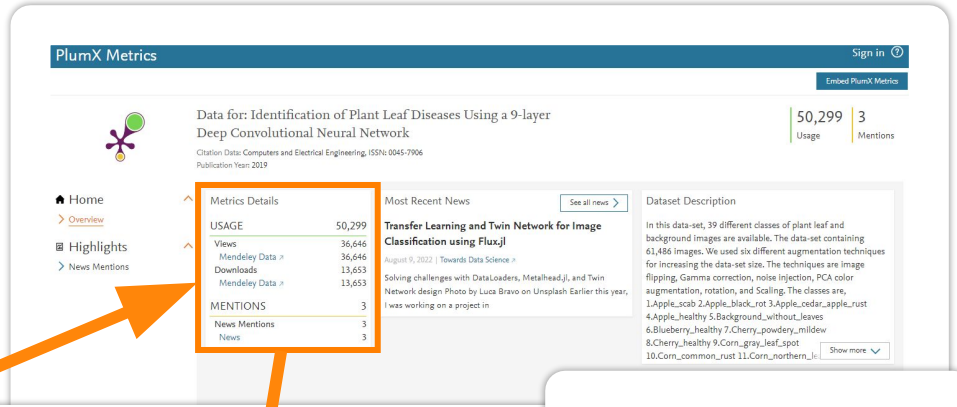


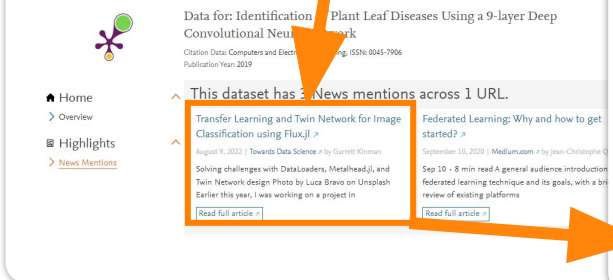
CENTER FOR OPEN SCIENCE

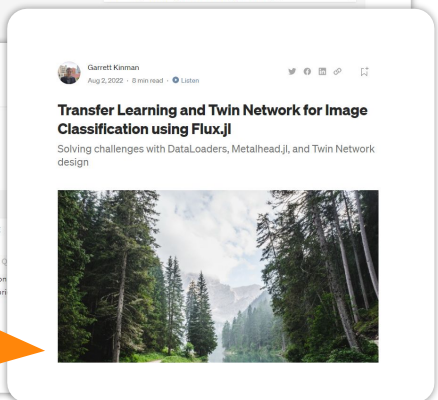
MENDELEY DATA for Data **Reuse**





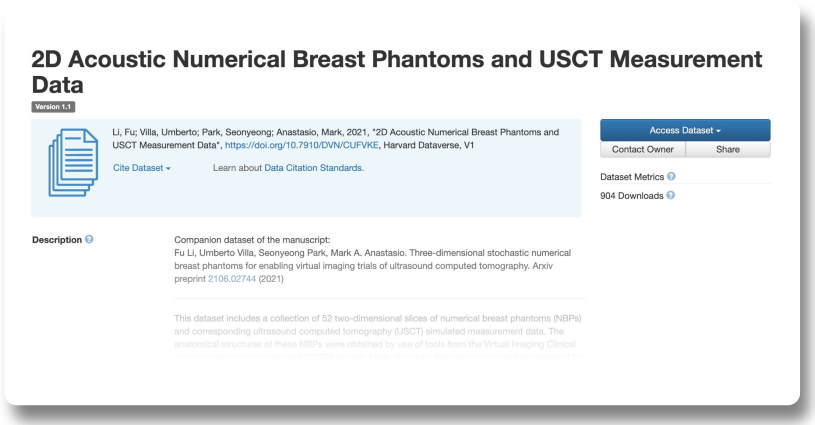




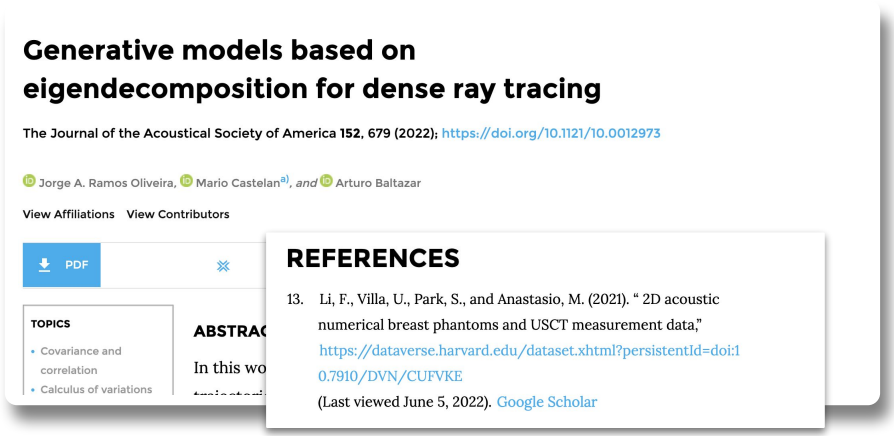


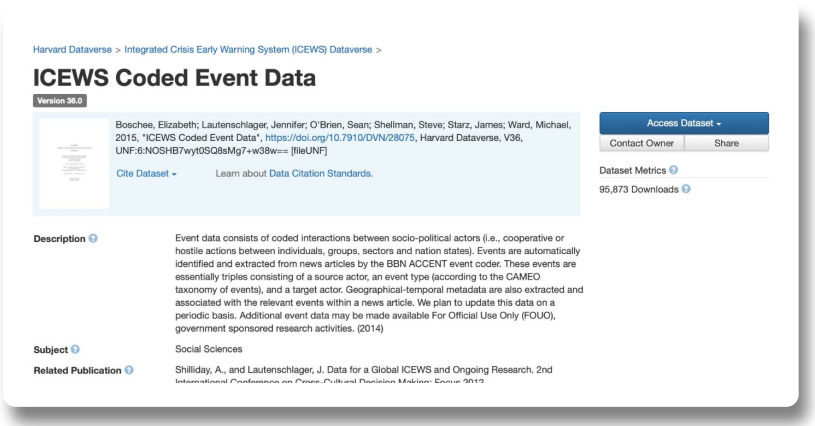
https://www.elsevier.com/rdm
https://data.mendeley.com/

Re-use of datasets published in HARVARD Dataverse



cited by





cited by

Data Metrics used by Generalists Repositories

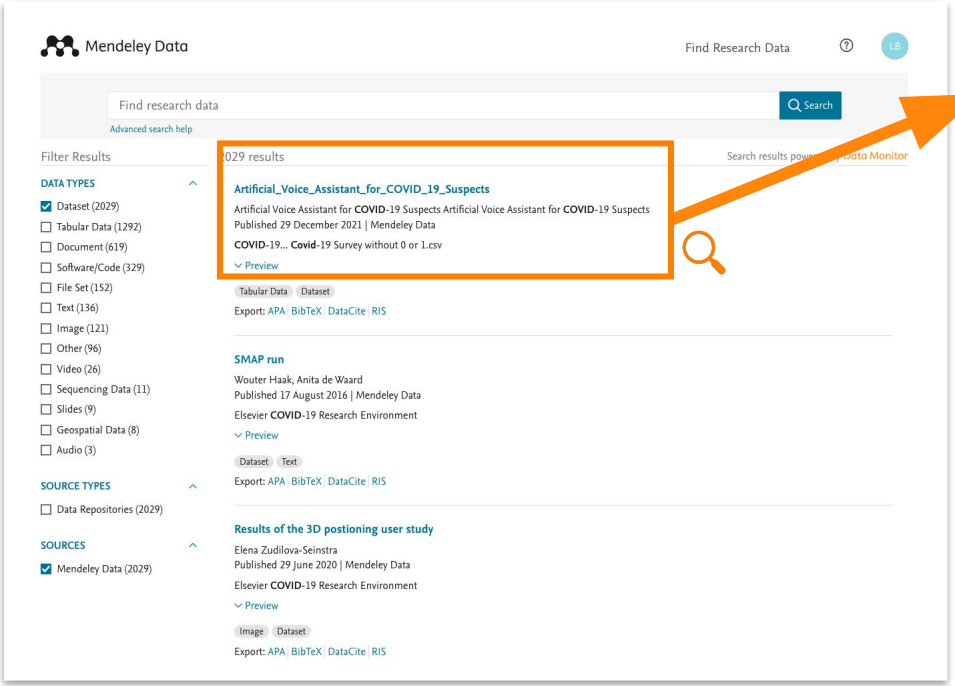
TOPIC	HARVARD DATAVERSE	DRYAD	FIGSHARE	MENDELEY DATA	OSF	VIVLI	ZENODO
Normalized data usage statistics (e.g. Make Data Count)	By end of Q2, will enable Counter Code of Practice for Reseach Data (Make Data Count)	Following standards: Counter Code of Practice for Reseach Data (Make Data Count) in both standardizing and reporting usage to DataCite	Currently enabling Counter Code of Practice for Reseach Data (Make Data Count)	Following standards: Counter Code Of Practice for Research Data Usage Metrics, starting Q3 2020 (Make Data Count)	No		Following standards: Counter Code Of Practice for Research Data Usage Metrics (Make Data Count)
Characteristics supporting Metrics							
Supported Data Use Metrics	Downloads, explorations, pageviews, data volume	Investigations (Views), Requests (Downloads), citations	View, downloads, citations, altmetrics	Views, downloads, altmetrics	Downloads (per version), Links, Forks	https://vivli.org/resources/platform_metrics-2/	Views, Downloads, Data Volume, Citations, Altmetrics

doi:10.5281/zenodo.3946720

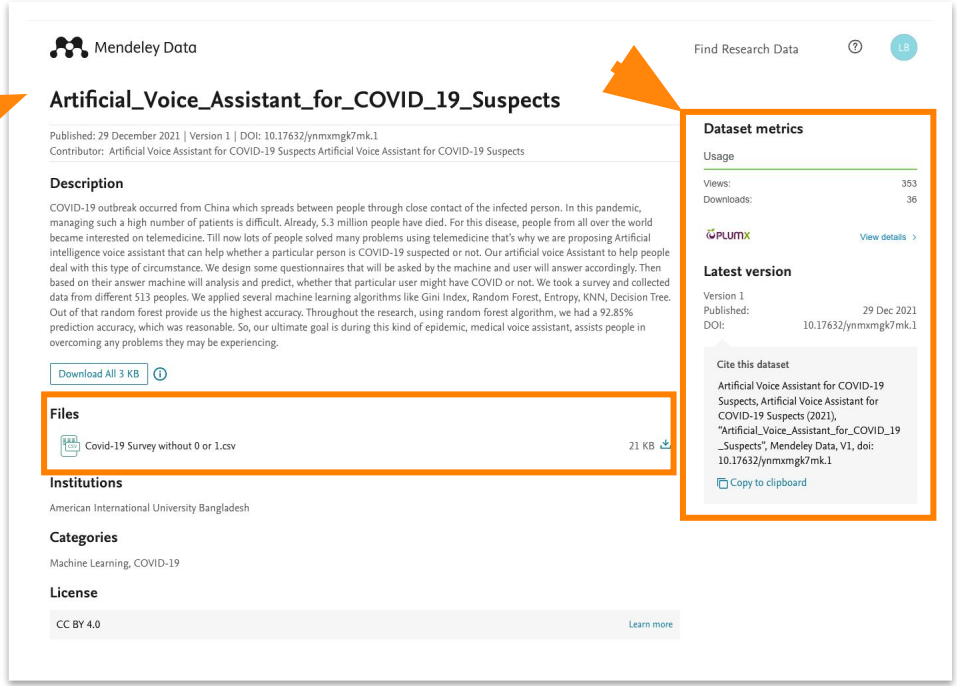
https://fairsharing.org/collection/GeneralRepositoryComparison

MENDELEY DATA for Data **Discovery**

Dataset Search Results



Dataset view



https://www.elsevier.com/rdm
https://data.mendeley.com/

Do you think you (or researchers you support) are more likely to share your data, knowing that you can see that your data is being used and cited?

Breakout Session

30 min

Here's the plan:

1. Everyone will randomly be assigned a 'room' to join.
2. Facilitators will ask questions for your **feedback** and/or you can pose your own **questions** for discussion.
3. After 30 minutes, we will regroup in the main meeting room.

Thank you for participating!
Have fun.

Breakout Session

30 min

Breakout session questions

1. How can generalist repositories help eliminate barriers to data discovery and reuse?
2. What challenges do you have with finding data?
3. What would enable more data reuse?

Breakout Session

30 min

Group 1:
Facilitator: Kelly Stathis, DataCite

Group 2:
Facilitator: Julie Wood, Vivli

Group 3:
Julian Gautier, Dataverse

Group 4:
Facilitator: Sara Gonzales, Zenodo

Group 5:
Facilitator: David Scherer, Elsevier

Group 6:
Facilitator: Eric Olson, OSF

Group 7:
Facilitator: Blaine Butler, OSF

GREI Workshop

Agenda

Tuesday, January 24	
11 – 11:05 a.m.	Welcome
11:05 – 11:15 a.m.	Introduction to GREI and Workshop Logistics
11:15 – 11:25 a.m.	Welcome from ODSS, GREI Program
11:25 a.m. – 12:20 p.m.	Day 1 Keynote
12:20 – 12:30 p.m.	Break
12:30 – 2 p.m.	Panel Session: Research Community Perspectives on Data Sharing
2 – 2:15 p.m.	Break
2:15 – 3:45 p.m.	Day 1 Interactive Training Session: Using generalist repositories to share data - exploring specific use cases and repository functionality
3:45 – 4 p.m.	Day 1 Wrap-up

Wednesday, January 25	
11 – 11:05 a.m.	Day 2 Welcome
11:05 a.m. – 12:45 p.m.	Panel Session: NIH stakeholder perspectives on generalist repositories in the data sharing landscape
12:45 – 1 p.m.	Break
1 – 1:50 p.m.	Day 2 Keynote
1:50 – 2 p.m.	Break
2 – 3:20 p.m.	Day 2 Interactive training session: Discovering and reusing data in generalist repositories
3 – 3:30 p.m.	Closing

Times are noted in Eastern Standard Time

https://datascience.nih.gov/news/grei-workshop-january-24-25-2023