



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)



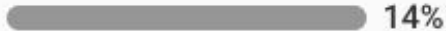
POLL

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

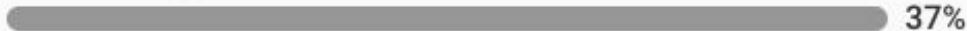


Have you shared data or supported a researcher in sharing data?

Yes, I have shared data



Yes, I have supported a researcher share data



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



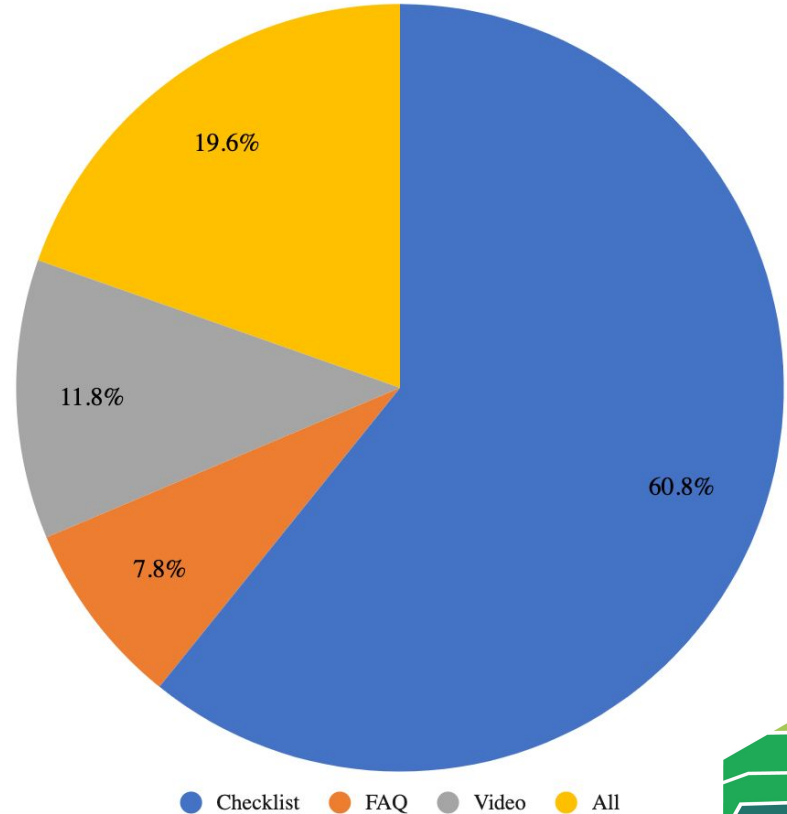
No, I have not shared data or supported a researcher with data sharing



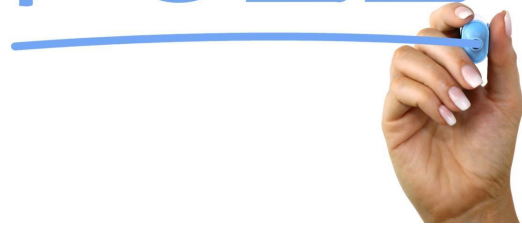
POLL



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?



POLL



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?

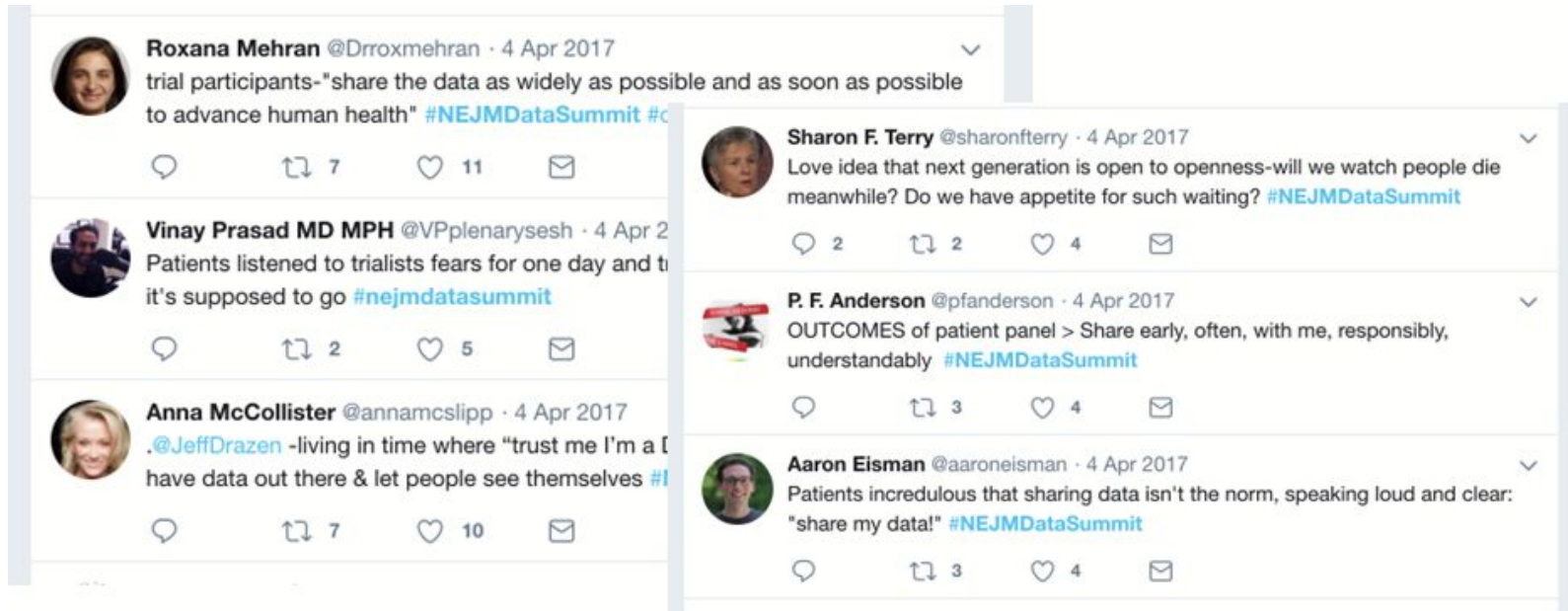
- High levels of support for data sharing; however, patients are reluctant to have their data “commodified” purely for commercial gain
- If adequate safeguards were in place, trial participants are willing to share their data

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



A screenshot of a Twitter thread from the #NEJMDataSummit. The tweets are arranged in two columns. The left column contains three tweets, and the right column contains three tweets. Each tweet includes a profile picture, the user's name and handle, the date (4 Apr 2017), and engagement metrics (replies, retweets, likes, and a share icon). The tweets discuss the importance of data sharing and reuse in clinical trials.

Roxana Mehran @Drroxmehrnan · 4 Apr 2017
trial participants-"share the data as widely as possible and as soon as possible to advance human health" [#NEJMDataSummit](#)

Vinay Prasad MD MPH @VPplenarysesh · 4 Apr 2017
Patients listened to trialists fears for one day and then it's supposed to go [#nejmdatsummit](#)

Anna McCollister @annamcslipp · 4 Apr 2017
.@JeffDrazen -living in time where "trust me I'm a doctor" have data out there & let people see themselves [#NEJMDataSummit](#)

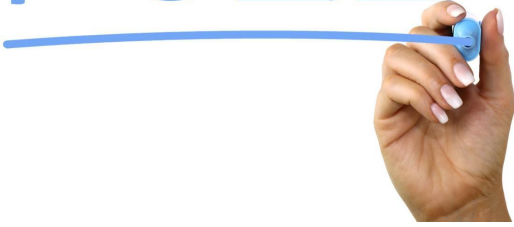
Sharon F. Terry @sharonferry · 4 Apr 2017
Love idea that next generation is open to openness-will we watch people die meanwhile? Do we have appetite for such waiting? [#NEJMDataSummit](#)

P. F. Anderson @pfanderson · 4 Apr 2017
OUTCOMES of patient panel > Share early, often, with me, responsibly, understandably [#NEJMDataSummit](#)

Aaron Eisman @aaroneisman · 4 Apr 2017
Patients incredulous that sharing data isn't the norm, speaking loud and clear: "share my data!" [#NEJMDataSummit](#)



POLL



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)



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>

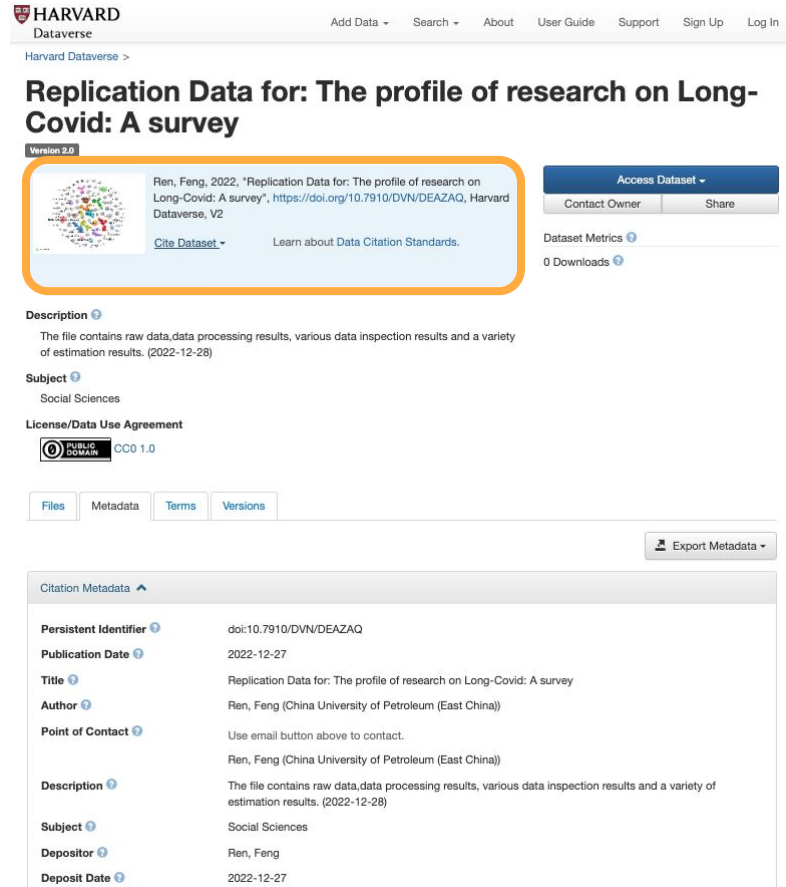


Example: Dataset DOI

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



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



The screenshot shows the Harvard Dataverse interface for a dataset. At the top, the Harvard Dataverse logo and navigation links (Add Data, Search, About, User Guide, Support, Sign Up, Log In) are visible. The dataset title is "Replication Data for: The profile of research on Long-Covid: A survey". Below the title, there is a "Version 2.0" label and a highlighted box containing a thumbnail image of a globe, the citation text "Ren, Feng, 2022, 'Replication Data for: The profile of research on Long-Covid: A survey', https://doi.org/10.7910/DVN/DEAZAQ, Harvard Dataverse, V2", and a "Cite Dataset" button. To the right of this box are buttons for "Access Dataset", "Contact Owner", and "Share". Below the highlighted box, there are sections for "Description" (The file contains raw data, data processing results, various data inspection results and a variety of estimation results. (2022-12-28)), "Subject" (Social Sciences), and "License/Data Use Agreement" (CC0 1.0). At the bottom, there are tabs for "Files", "Metadata", "Terms", and "Versions", and an "Export Metadata" button. The "Citation Metadata" section is expanded, showing fields like Persistent Identifier, Publication Date, Title, Author, Point of Contact, Description, Subject, Depositor, and Deposit Date.



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 (CCo).

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:

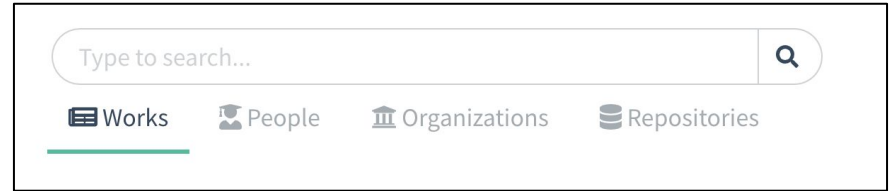
- Works (DOIs)
- Researchers
- Organizations
- Repositories
- Citations, views, downloads and more...


And they are all connected through links in the metadata.







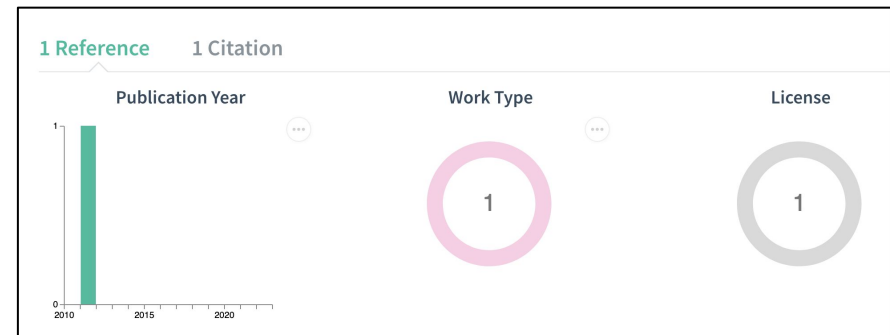
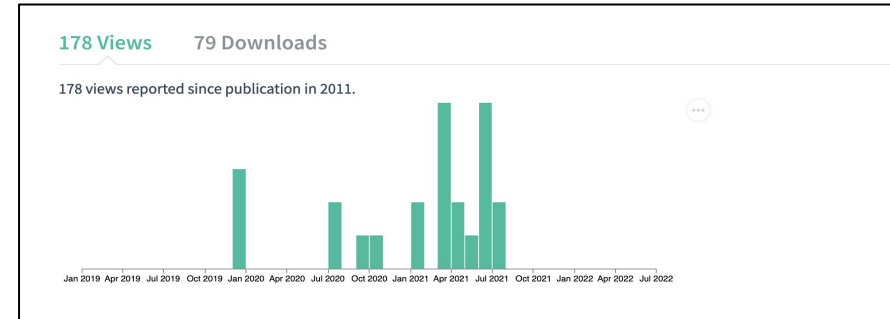
DataCite Commons

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



Type to search... 

 Works  People  Organizations  Repositories



Example - search for datasets

DataCite Commons

tuberculosis

Pages Support

4 Works

Accuracy of the InnovaveDX MTB/RIF test for detection of *Mycobacterium tuberculosis* and rifampicin resistance: a prospective multicentre study

Yunfeng Deng, Zichun Ma, Biyi Su, Guanghong Bai, Jianhua Pan, Quan Wang, Long Cai, Yanhua Song, Yuanyuan Shang, Pinyun Ma, Jing Li, Qianxuan Zhou, Gulibike Mulati, Dapeng Fan, Shanshan Li, Yaoju Tan & Yu Pang
Dataset published 2023 in [figshare Academic Research System](#)

Early and accurate diagnosis of tuberculosis (TB) is necessary to initiate proper therapy for the benefit of the patients and to prevent disease transmission in the community. In this study, we developed the InnovaveDX MTB/RIF (InnovaveDX) to detect *Mycobacterium tuberculosis* (MTB) and rifampicin resistance simultaneously. A prospective multicentre study was conducted to evaluate the diagnostic performance of InnovaveDX for the detection MTB in sputum samples as compared with Xpert and culture. The calculated limit of detection (LOD) for InnovaveDX was 9.6 CFU/ml for TB detection and 374.9 CFU/ml for RIF susceptibility. None of the other bacteria tested produced signals that fulfilled the positive TB criteria, demonstrating a species-specificity of InnovaveDX. Then 951 individuals were enrolled at 7 hospitals, of which 607 were definite TB cases with positive culture and/or Xpert results, including 354 smear-positive and 253 smear-negative cases. InnovaveDX sensitivity was 92.7% versus bacteriologically TB standard. Further follow-up revealed that 61 (91.0%) out of 67 false-positive patients with no bacteriological evidence met the criteria of clinically diagnosed TB. Among 125 RIF-resistant TB patients diagnosed by Xpert, 108 cases were correctly identified by InnovaveDX, yielding a sensitivity of 86.4%. Additionally, the proportion of very low bacterial load in the discordant susceptibility group was significantly higher than in the concordant susceptibility group ($P = 0.029$). To conclude, we have developed a novel molecular diagnostic with promising detection capabilities of TB and RIF susceptibility. In addition, the discordant RIF susceptibility results between InnovaveDX and Xpert are more frequently observed in samples with very low bacterial load.

DOI registered January 2, 2023 via DataCite.



<https://doi.org/10.6084/m9.figshare.21804085>

Publication Year

2023 4

Work Type

Dataset 4

License

CC-BY-4.0 4

Field of Science

Biological sciences 4

Health sciences 4

Chemical sciences 2

Mathematics 2

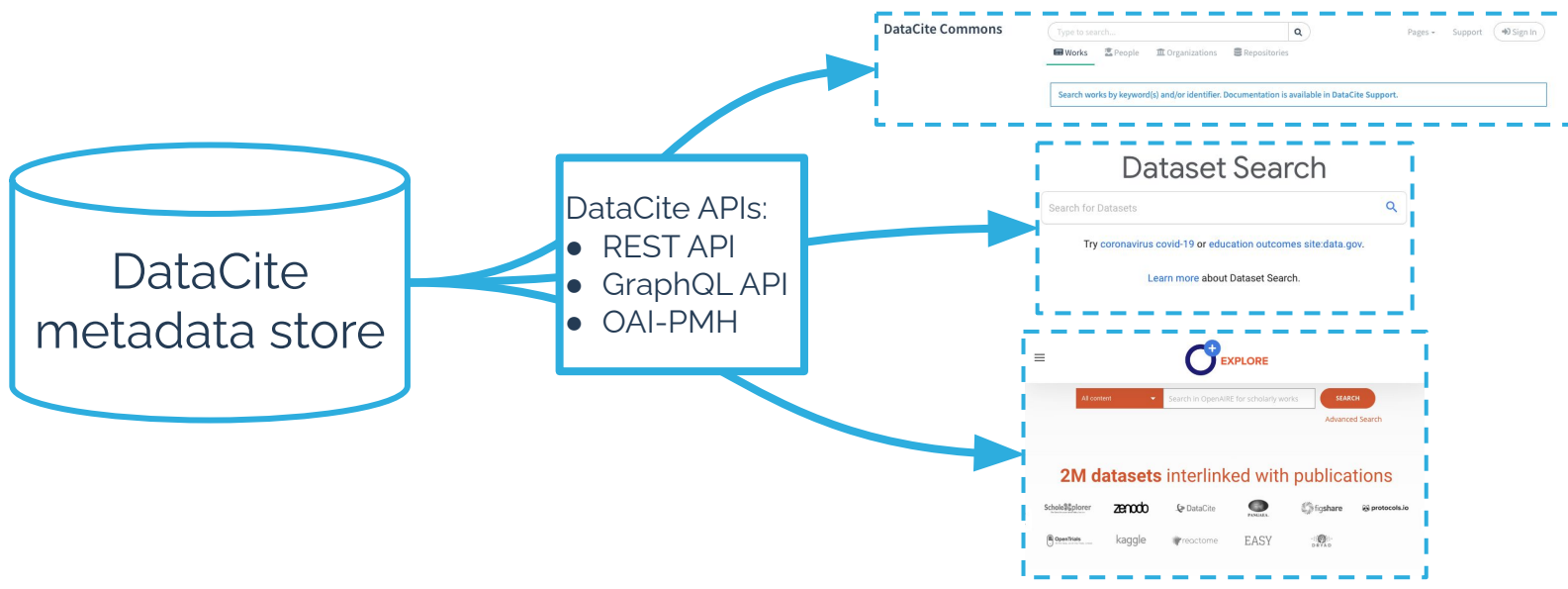
Registration Agency

DataCite 4

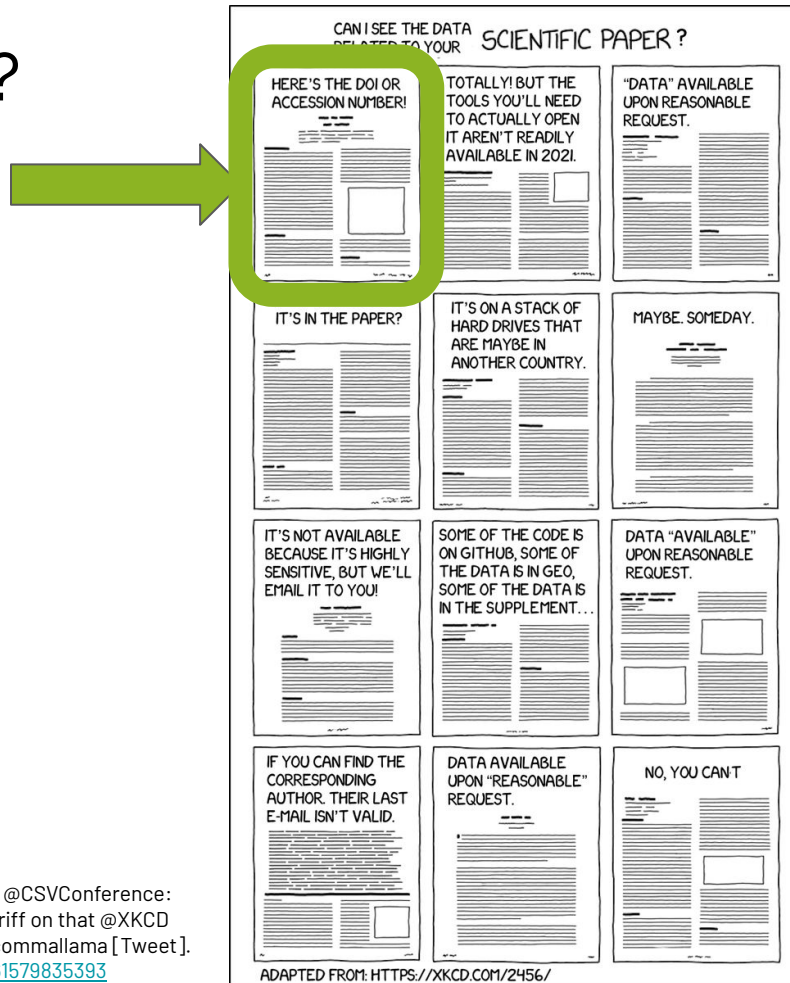


Harvesters and aggregators

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

Cites

DRYAD

Search

Explore data | About | Help | Login

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

Schulte, Georg
Hübner, Ursula
Rienhoff, Otto
Quade, Matthias
Rottmann, Thorsten
Fenske, Matthias
Egbert, Nicole
Kuhlisch, Raik
Sellemann, Björn
Publication date: October 16, 2018
Publisher: Dryad
<https://doi.org/10.5061/dryad.9f2d8>

Citation

Schulte, Georg et al. (2018). Data from: Evaluation of electronically supported nursing transfers between hospital and nursing home based on a test health telematics infrastructure: a case analysis, Dryad, Dataset, <https://doi.org/10.5061/dryad.9f2d8>

Works referencing this dataset

Schulte, Georg et al. (2017), Evaluation einer elektronisch unterstützten pflegerischen Überleitung zwischen Krankenhaus und Pflegeheim unter Nutzung einer Test-Telematikinfrastruktur: eine Fallanalyse, GMS Medizinische Informatik, Article-journal, <https://doi.org/10.3205/mibe000172>

224 views
51 downloads
1 citations

Is Cited By

GMS German Medical Science

Deutsch | MBE | About MBE | For authors | Contact | Imprint | GMS Meetings

Portal
Journals
Meetings
Reports
Guidelines
Handbooks

gmds | GMS Medizinische Informatik, Biometrie und Epidemiologie
Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (DGMB) | ISSN 1860-9171

Article | Current Volume | Archive | Search in MBE | Newsletter

Originalarbeit

Evaluation einer elektronisch unterstützten pflegerischen Überleitung zwischen Krankenhaus und Pflegeheim unter Nutzung einer Test-Telematikinfrastruktur: eine Fallanalyse

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

Georg Schulte - Forschungsgruppe Informatik im Gesundheitswesen, Hochschule Osnabrück, Osnabrück, Deutschland; Klinikum Osnabrück GmbH, Osnabrück, Deutschland
Ursula Hübner - Forschungsgruppe Informatik im Gesundheitswesen, Hochschule Osnabrück, Osnabrück, Deutschland
Otto Rienhoff - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland
Matthias Quade - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland
Thorsten Rottmann - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland
Matthias Fenske - Dakonwerk Osnabrück gGmbH, Osnabrück, Deutschland
Nicole Egbert - Forschungsgruppe Informatik im Gesundheitswesen, Hochschule Osnabrück, Osnabrück, Deutschland
Raik Kuhlisch - Fraunhofer FOKUS, Berlin, Deutschland
Björn Sellemann - Institut für Medizinische Informatik, Universitätsmedizin Göttingen, Göttingen, Deutschland; Interdisziplinäre Notfallaufnahme, Universitätsmedizin Göttingen, Göttingen, Deutschland

GMS Med Inform Biom Epidemiol 2017;17(1):Dec05
[doi: 10.3205/mibe000172](https://doi.org/10.3205/mibe000172) | [arxiv:1803.01833](https://arxiv.org/abs/1803.01833) | [mibe000172](https://doi.org/10.3205/mibe000172)



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 ePfgebericht 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 Döringer, Regina Thalhammer, Katharina Lüftl, Birgit Prodingler & Martin Müller
Journal Article published 2023 in *BMC Nursing*

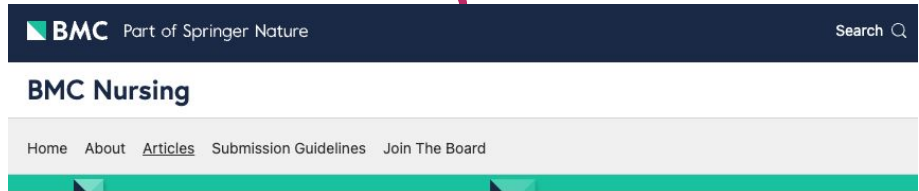
Other Identifiers
article-number: 5

DOI registered January 8, 2023 via Crossref.



Journal Article

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



BMC Part of Springer Nature

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 Prodingler](#) & [Martin Müller](#)

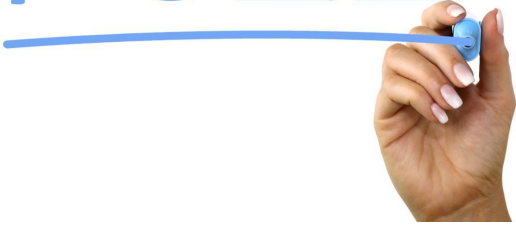
BMC Nursing 22, Article number: 5 (2023) | [Cite this article](#)

270 Accesses | [Metrics](#)

- 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



POLL



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 Mingqing, 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., Mingqing, 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/tcon.ggg2014.ascension01.ro/1149285>



Data Re-use Case study



The screenshot shows the Vivli website header with navigation links: Home, About, Members, News & Events, Resources, Find Studies, and MY ACCOUNT. The Vivli logo is on the left, and social media icons for Twitter, Facebook, and LinkedIn are on the right. Below the header is a dark blue banner with white text that reads: "A systematic review and individual patient data meta-analysis of physiological biomarkers in idiopathic pulmonary fibrosis".

Lead Investigator: Fasihul Khan, University of Nottingham

Title of Proposal Research: A systematic review and individual patient data meta-analysis of physiological biomarkers in idiopathic pulmonary fibrosis

Vivli Data Request: 5207

Funding Source: PI is funded by NIHR grant

Conflicts of Interest: RGJ reports grants from GlaxoSmithKline, UK Medical Research Council, Biogen, Galecto, MedImmune; as well as personal fees from Boehringer Ingelheim, Galapagos, GlaxoSmithKline, Heptares, MedImmune, Roche and Pulmatrix; served as consultant for NuMedii and Pliant; a trustee for charities Action for Pulmonary Fibrosis and the British Thoracic Society

None of these will impact data analysis and publication.

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.

The screenshot shows the OSF Registries search results page. At the top, there is a search bar with the text "Search registrations..." and a question mark icon. Below the search bar, the page displays "1,148 Registrations" and a "Sort by: Relevance" dropdown menu. The main content area is divided into three sections, each with a title, authors, and a brief description. The first section is titled "Demographic change and shifting group boundaries in Germany: The effect of group threat on perceptions of who has a migration background" by Johanna Gereke, Joshua Heliyer, Susanne Veit, Nan Zhang, Jan Behnert, Alexander Herbel, Felix Jäger, Dean Lajic, Štěpán Mezenský, Vu Ngoc Anh, Tymoteusz Oglaza, Jule Schabinger, Anna Sokolova, Szafran, Daria, Noah Tirolf. The second section is titled "Replicating the original Stroop study (1935) - Fruit Images vs. Fruit Names" by Gladys Yeung, Crystal Yuen, Yue Hin Dominic Chan, Tsz Lok Fong, Victoria Sze, Tse Hiu Ching, Wendy Lau, Henry K S Ng. The third section is titled "Corruption and Hierarchy: A Replication of Studies 1c and 6 of Fath & Kay 2018". Each section includes an "Open resources" button and a list of resource types: Data, Analytic code, Materials, Papers, and Supplements. On the left side, there is a "Refine Search" section with a filter for "OSF Registries" and a list of providers with their respective registration counts. Below that is an "OSF Registration Type" section with a list of registration types and their counts.

OSF Registries

Search registrations...

1,148 Registrations

Sort by: Relevance

Demographic change and shifting group boundaries in Germany: The effect of group threat on perceptions of who has a migration background

Johanna Gereke, Joshua Heliyer, Susanne Veit, Nan Zhang, Jan Behnert, Alexander Herbel, Felix Jäger, Dean Lajic, Štěpán Mezenský, Vu Ngoc Anh, Tymoteusz Oglaza, Jule Schabinger, Anna Sokolova, Szafran, Daria, Noah Tirolf

OSF Registries | Replication Recipe (Brandt et al., 2013): Pre-Registration

A replication and extension of Abascal 2020 in the German context, examining national classifications under group threat cond...

Open resources

- Data
- Analytic code
- Materials
- Papers
- Supplements

Replicating the original Stroop study (1935) - Fruit Images vs. Fruit Names

Gladys Yeung, Crystal Yuen, Yue Hin Dominic Chan, Tsz Lok Fong, Victoria Sze, Tse Hiu Ching, Wendy Lau, Henry K S Ng

OSF Registries | Replication Recipe (Brandt et al., 2013): Pre-Registration

The study aims to replicate the original Stroop study conducted in 1935.

Open resources

- Data
- Analytic code
- Materials
- Papers
- Supplements

Data & Analysis

Thomas Rhys Evans, Susannah O'Regan, Renata Kviatkovskytė, Floriانا O Nkagbu Chukwudi, Nishat Tasnim, Shernay A Adolph

OSF Registries | Replication Recipe (Brandt et al., 2013): Pre-Registration

Corruption and Hierarchy: A Replication of Studies 1c and 6 of Fath & Kay 2018

Open resources

Refine Search

OSF Registries x

Replication Recipe (Brandt et al., 2013): Pre-Registration x

Provider

<input type="checkbox"/> ClinicalTrials.gov	366,269
<input type="checkbox"/> Research Registry	3,405
<input type="checkbox"/> Character Lab Registry	416
<input type="checkbox"/> DARPA ASIST Registry	60
<input type="checkbox"/> egap Registry	2,591
<input type="checkbox"/> Metascience Registry	6
<input checked="" type="checkbox"/> OSF Registries	118,191
<input type="checkbox"/> Real World Evidence Registry	45
<input type="checkbox"/> YOUTH Study Registry	10

OSF Registration Type

Only available with OSF Registries

<input type="checkbox"/> Election Research Preacceptance Competition	
<input type="checkbox"/> OSF Preregistration	
<input type="checkbox"/> OSF-Standard Pre-Data Collection Registration	
<input type="checkbox"/> Open-Ended Registration	
<input type="checkbox"/> Pre-Registration in Social Psychology (van 't Veer & Giner-Sorolla, 2016): Pre-Registration	
<input type="checkbox"/> Preregistration Template from AsPredicted.org	
<input type="checkbox"/> Qualitative Preregistration	
<input type="checkbox"/> Registered Report Protocol Preregistration	
<input type="checkbox"/> Replication Recipe (Brandt et al., 2013): Post-Completion	
<input checked="" type="checkbox"/> Replication Recipe (Brandt et al., 2013): Pre-Registration	
<input type="checkbox"/> Secondary Data Preregistration	



Re-use on the OSF

Systemizing Confidence in Open Research and Evidence(SCORE)

Initial author makes data publicly available
2017

Data is then used to determine reproducibility claims
2021

The screenshot shows the OSF project page for "A Cognitive-Ecological Model of Intergroup Bias". The project title is highlighted in an orange box. Below the title, it lists the contributor as Hans Alves, with a creation date of 2017-10-05 and a last update of 2022-05-09. The "Files" section shows a folder named "A Cognitive-Ecological Model of Intergroup Bias" containing "OSF Storage (United States)", "data", and "Supplemental Materials". The "Supplemental Materials" folder contains a file named "Supplemental Materials.docx" with a modification date of 2018-01-05. The "Citation" and "Recent Activity" sections are also visible.

482.6KB Public P 0

The screenshot shows the OSF project page for "Alves_PsychologSci_2018_AvOr - Hanel". The project title is highlighted in an orange box. Below the title, it lists contributors: Elan Simon Parsons, Olivia Miske, Bri Luis [SCORE], Zachary Loomas, Nicholas Fox [SCORE], Andrew Tynner, Krystal Hahn, Paul H.P. Hanel. The creation date is 2021-11-30 and the last update is 2023-01-20. The "Wiki" section contains text about file public access and a link to "VIEWONLY_MATERIALS_LINK". The "Files" section shows a folder named "Alves_PsychologSci_2018_AvOr - Hanel - 21352" containing files like "Alves_PsychologSci_2018_AvOr_21352_PBR.xlsx" and "Alves_PsychologSci_2018_AvOr_busthel_claims.md".

<https://osf.io/qenhu>



CENTER FOR
OPEN SCIENCE



Filter

TensorFlow > Resources > Datasets > Catalog

plant_village

Visualization: Explore in Know Your Data

Description:

The PlantVillage dataset consists of 54303 healthy and unhealthy leaf images divided into 38 categories by species and disease.

Note: The original dataset is not available from the original source (plantvillage.org), therefore we get the unaugmented dataset from a paper that used that dataset and republished it. Moreover, we dropped images with Background_without_leaves label, because these were not present in the original dataset.

Original paper URL: <https://arxiv.org/abs/1511.08960> Dataset URL: <https://data.mendeley.com/datasets/ywbtspjv/1>



MENDELEY DATA for Data Reuse

Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network

Published: 17 April 2019 | Version 1 | DOI: 10.17632/ywbtspjv.1
Contributors: ARUN PANDIAN J., GEETHARAMANI GOPAL

Description

In this data-set, 39 different classes of plant leaf and background images are available. The data-set containing 61,486 images. We used six different augmentation techniques for increasing the data-set size. The techniques are image flipping, Gamma correction, noise injection, PCA color augmentation, rotation, and Scaling.

- The classes are,
- 1.Apple_scab
 - 2.Apple_black_rot
 - 3.Apple_cedar_apple_rust
 - 4.Apple_healthy
 - 5.Background_without_leaves
 - 6.Blueberry_healthy
 - 7.Cherry_powdery_mildew
 - 8.Cherry_healthy
 - 9.Corn_gray_Leaf_spot
 - 10.Corn_common_rust
 - 11.Corn_northern_Leaf_blight
 - 12.Corn_healthy
 - 13.Grape_black_rot

9 Citations

Dataset metrics

Usage	
Views	36,646
Downloads	13,653
Mentions	
News Mentions	3

Latest version

Version 1
Published: 17 Apr 2019
DOI: 10.17632/ywbtspjv.1

Cite this dataset

PlumX Metrics

Sign in

Embed PlumX Metrics

50,299 Usage | 3 Mentions

Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network

Citation Data: Computers and Electrical Engineering, ISSN: 0045-7906
Publication Year: 2019

Metrics Details	
USAGE	50,299
Views	36,646
Mendeley Data >	36,646
Downloads	13,653
Mendeley Data >	13,653
MENTIONS	3
News Mentions	3
News	3

Most Recent News

Transfer Learning and Twin Network for Image Classification using Flux.jl

Aug 9, 2022 | Towards Data Science >

Solving challenges with DataLoaders, Metalhead.jl, and Twin Network design Photo by Luca Bravo on Unsplash Earlier this year, I was working on a project in

Dataset Description

In this data-set, 39 different classes of plant leaf and background images are available. The data-set containing 61,486 images. We used six different augmentation techniques for increasing the data-set size. The techniques are image flipping, Gamma correction, noise injection, PCA color augmentation, rotation, and Scaling. The classes are, 1.Apple_scab 2.Apple_black_rot 3.Apple_cedar_apple_rust 4.Apple_healthy 5.Background_without_leaves 6.Blueberry_healthy 7.Cherry_powdery_mildew 8.Cherry_healthy 9.Corn_gray_Leaf_spot 10.Corn_common_rust 11.Corn_northern_Leaf_blight 12.Corn_healthy 13.Grape_black_rot

Data for: Identification of Plant Leaf Diseases Using a 9-layer Deep Convolutional Neural Network

Citation Data: Computers and Electrical Engineering, ISSN: 0045-7906
Publication Year: 2019

This dataset has 3 News mentions across 1 URL.

Transfer Learning and Twin Network for Image Classification using Flux.jl >

Aug 9, 2022 | Towards Data Science > by Garrett Kinman

Solving challenges with DataLoaders, Metalhead.jl, and Twin Network design Photo by Luca Bravo on Unsplash Earlier this year, I was working on a project in

Federated Learning: Why and how to get started? >

September 10, 2020 | Medium.com > by Jean-Christophe Q

Sep 10 · 8 min read A general audience introduction to federated learning technique and its goals, with a brief review of existing platforms

Read full article >

Garrett Kinman
Aug 2, 2022 · 6 min read · Listen

Transfer Learning and Twin Network for Image Classification using Flux.jl

Solving challenges with DataLoaders, Metalhead.jl, and Twin Network design

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



Re-use of datasets published in HARVARD Dataverse

2D Acoustic Numerical Breast Phantoms and USCT Measurement Data

Version 1.1



Li, Fu; Villa, Umberto; Park, Seonyeong; Anastasio, Mark, 2021, "2D Acoustic Numerical Breast Phantoms and USCT Measurement Data", <https://doi.org/10.7910/DVN/CUFVKE>, Harvard Dataverse, V1

Cite Dataset - Learn about Data Citation Standards.

Access Dataset -
Contact Owner Share

Dataset Metrics

904 Downloads

Description

Companion dataset of the manuscript:

Fu Li, Umberto Villa, Seonyeong Park, Mark A. Anastasio. Three-dimensional stochastic numerical breast phantoms for enabling virtual imaging trials of ultrasound computed tomography. *ArXiv preprint 2106.02744* (2021)

This dataset includes a collection of 52 two-dimensional slices of numerical breast phantoms (NBPs) and corresponding ultrasound computed tomography (USCT) simulated measurement data. The anatomical structures of these NBPs were obtained by use of tools from the Virtual Imaging Clinical Applications (VICA) software suite. This data is available for the public domain and for research purposes.

cited by



Generative models based on eigendecomposition for dense ray tracing

The Journal of the Acoustical Society of America 152, 679 (2022); <https://doi.org/10.1121/10.0012973>

Jorge A. Ramos Oliveira, Mario Castelan¹⁾, and Arturo Baltazar

View Affiliations View Contributors

PDF

TOPICS

- Covariance and correlation
- Calculus of variations

ABSTRACT

In this wo

REFERENCES

- Li, F., Villa, U., Park, S., and Anastasio, M. (2021). "2D acoustic numerical breast phantoms and USCT measurement data," <https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/CUFVKE> (Last viewed June 5, 2022). [Google Scholar](#)

Harvard Dataverse > Integrated Crisis Early Warning System (ICEWS) Dataverse >

ICEWS Coded Event Data

Version 26.0



Boschee, Elizabeth; Lautenschlager, Jennifer; O'Brien, Sean; Shellman, Steve; Starz, James; Ward, Michael, 2015, "ICEWS Coded Event Data", <https://doi.org/10.7910/DVN/28075>, Harvard Dataverse, V36, UNF:6:NOSHb7wyt0SQ8sMg7+w38w== [fileUNF]

Cite Dataset - Learn about Data Citation Standards.

Access Dataset -
Contact Owner Share

Dataset Metrics

95,873 Downloads

Description

Event data consists of coded interactions between socio-political actors (i.e., cooperative or hostile actions between individuals, groups, sectors and nation states). Events are automatically identified and extracted from news articles by the BBN ACCENT event coder. These events are essentially triples consisting of a source actor, an event type (according to the CAMEO taxonomy of events), and a target actor. Geographical-temporal metadata are also extracted and associated with the relevant events within a news article. We plan to update this data on a periodic basis. Additional event data may be made available For Official Use Only (FOUO), government sponsored research activities. (2014)

Subject

Social Sciences

Related Publication

Shilliday, A., and Lautenschlager, J. Data for a Global ICEWS and Ongoing Research. 2nd International Conference on Peace, Cultural Politics, Politics, Events 2019.

cited by



Journal of Asian Economics

Volume 84, February 2023, 101578



The Arab Spring, a setback for gender equality? Evidence from the Gallup World Poll

Robert Rudolf* ✉, Sh

Show more

+ Add to Mendeley

References

- Boschee et al., 2015 Boschee, E., Lautenschlager, J., O'Brien, S., Shellman (2015). ICEWS coded event data. Harvard Dataverse, V28, UNF:6:NOSHb7wyt0SQ8sMg7+w38w== [fileUNF]. Retrieved from <<https://doi.org/10.7910/DVN/28075>> . [Google Scholar](#)



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 Research Data (Make Data Count)	Following standards: Counter Code of Practice for Research Data (Make Data Count) in both standardizing and reporting usage to DataCite	Currently enabling Counter Code of Practice for Research 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

4

<https://fairsharing.org/collection/GeneralRepositoryComparison>



MENDELEY DATA for Data Discovery

Dataset Search Results

Dataset view

Mendeley Data Find Research Data

Find research data Search

Advanced search help

Filter Results

209 results

Artificial_Voice_Assistant_for_COVID_19_Suspects
Artificial Voice Assistant for COVID-19 Suspects Artificial Voice Assistant for COVID-19 Suspects
Published 29 December 2021 | Mendeley Data
COVID-19... Covid-19 Survey without 0 or 1.csv
Preview

Tabular Data Dataset

Export: APA BibTeX DataCite RIS

SMAP run
Wouter Haak, Anita de Waard
Published 17 August 2016 | Mendeley Data
Elsevier COVID-19 Research Environment
Preview

Dataset Text

Export: APA BibTeX DataCite RIS

Results of the 3D positioning user study
Elena Zudilova-Seinstra
Published 29 June 2020 | Mendeley Data
Elsevier COVID-19 Research Environment
Preview

Image Dataset

Export: APA BibTeX DataCite RIS

Mendeley Data Find Research Data

Artificial_Voice_Assistant_for_COVID_19_Suspects

Published: 29 December 2021 | Version 1 | DOI: 10.17632/jnmxmgk7mk.1
Contributor: Artificial Voice Assistant for COVID-19 Suspects Artificial Voice Assistant for COVID-19 Suspects

Description
COVID-19 outbreak occurred from China which spreads between people through close contact of the infected person. In this pandemic, managing such a high number of patients is difficult. Already, 5.3 million people have died. For this disease, people from all over the world become interested in telemedicine. Till now lots of people solved many problems using telemedicine that's why we are proposing Artificial Intelligence voice assistant that can help whether a particular person is COVID-19 suspected or not. Our artificial voice Assistant to help people deal with this type of circumstance. We design some questionnaires that will be asked by the machine and user will answer accordingly. Then based on their answer machine will analyze and predict, whether that particular user might have COVID or not. We took a survey and collected data from different 513 peoples. We applied several machine learning algorithms like Gini Index, Random Forest, Entropy, KNN, Decision Tree. Out of that random forest provides us the highest accuracy. Throughout the research, using random forest algorithm, we had a 92.85% prediction accuracy, which was reasonable. So, our ultimate goal is during this kind of epidemic, medical voice assistant, assists people in overcoming any problems they may be experiencing.

Download All 3 KB

Files
Covid-19 Survey without 0 or 1.csv 21 KB

Institutions
American International University Bangladesh

Categories
Machine Learning, COVID-19

License
CC BY 4.0 Learn more

Dataset metrics

Usage

Views:	353
Downloads:	36

PLUMX View details >

Latest version

Version 1	
Published:	29 Dec 2021
DOI:	10.17632/jnmxmgk7mk.1

Cite this dataset

Artificial Voice Assistant for COVID-19 Suspects, Artificial Voice Assistant for COVID-19 Suspects (2021), "Artificial_Voice_Assistant_for_COVID_19_Suspects", Mendeley Data, v1, doi: 10.17632/jnmxmgk7mk.1

Copy to clipboard

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



POLL



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>

