

Dr Yanti Idaya Aspura binti Mohd Khalid
Head of Department

Jabatan Sains Perpustakaan dan Maklumat | Fakulti Sastera & Sains Sosial
yanti@um.edu.my

Serving the Nation. Impacting the World.



UNIVERSITI
MALAYA

Research Data Management for Information Professional

Serving the Nation. Impacting the World.

www.um.edu.my



Content

Research Data

Research Data Management (RDM)

What does RDM offer to Information Professional?

Example

Xxxx

xxx

3

Research Data

“... that is collected, observed, or created, for purposes of analysis to produce original research results”

(UK Data Archive, n.d.)

Formats

- Text
- Numerical
- Multimedia
- Software
- Discipline specific
- Instruments specific

Research Data Objects: Physical

- Physical: Documents (text, Word), spreadsheets
- Laboratory notebooks, field notebooks, diaries
- Questionnaires, transcripts, codebooks
- Audiotapes, videotapes
- Photographs, films
- Test responses
- Slides, artefacts, specimens, sample

Research Data Objects: Digital

- Statistical or other data files
- Database contents (video, audio, text, images) Models, algorithms, scripts
- Contents of an application (input, output, logfiles for analysis software, simulation software, schemas)
- Methodologies and workflows
- Standard operating procedures and protocols
- Research records

Types of research data



Raw Data



Curated Data



Published Data



Metadata



Today's Research Data is tomorrow's library collection...

Research Data Management

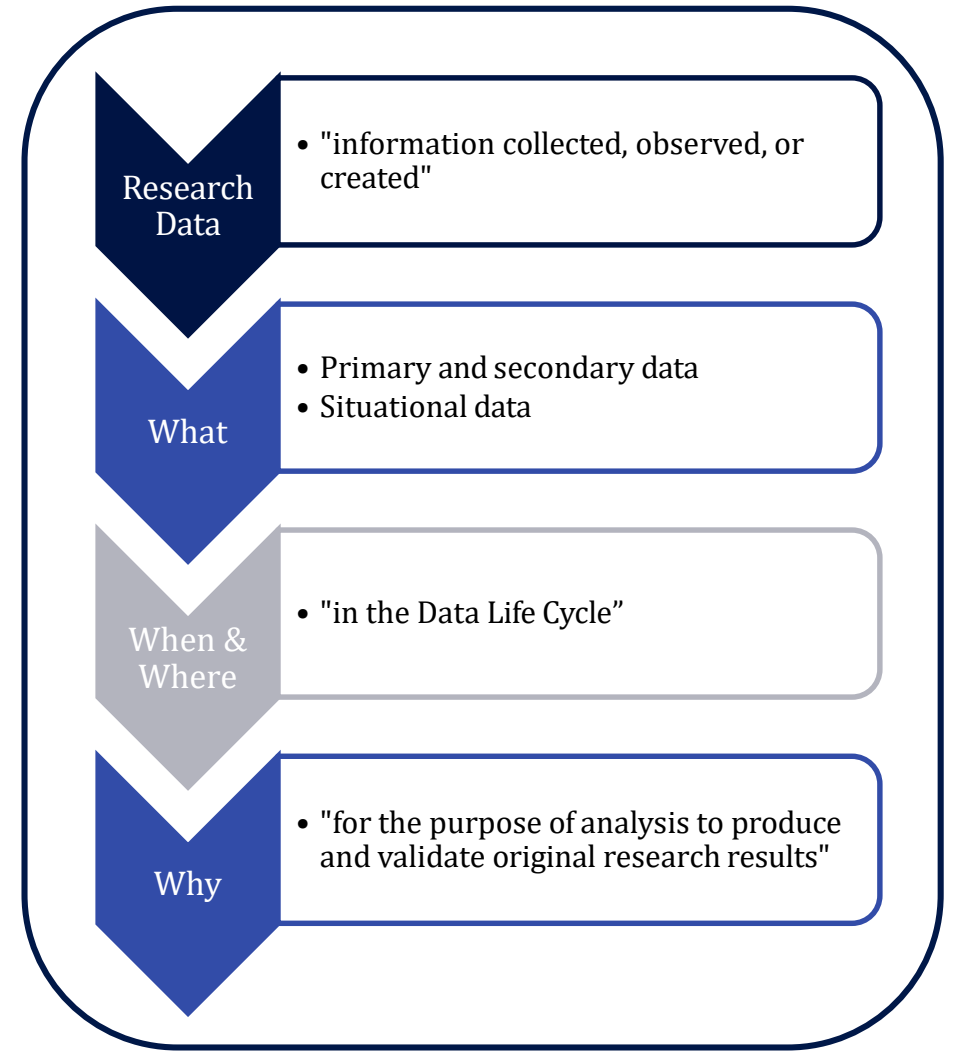
“ Research data management is “the active management and appraisal of data over the lifecycle of scholarly and scientific interest ”

Digital Curation Center, UK.

organisation, storage and preservation of data created during a research

involves the **active organization** and **maintenance** of data throughout the **research process**, and suitable **archiving** of the data at the project’s completion.

It is an **ongoing activity** throughout the **data lifecycle** project



What does RDM offer to information professionals?

Chance

- To establish **credibility** in a new area of engagement
- To learn new “**technical skills**”.

Opportunity

- To **get closer** to the research community and their **research processes** (closer working relationship)
- To **get your hands dirty** with unpublished or raw materials that are the **building blocks** of knowledge

Role & Engagement

Serving the Nation. Impacting the World.



www.um.edu.my



[universityofmalaya](https://www.facebook.com/universityofmalaya)



[unimalaya](https://www.instagram.com/unimalaya)



[uniofmalaya](https://www.youtube.com/uniofmalaya)



UNIVERSITI
MALAYA

What research data management processes are we supporting?

Need to support and suppose to support

Serving the Nation. Impacting the World.

www.um.edu.my



Supporting



RESEARCH WORKFLOW

DMPs, existing data,
documentation, datasets,
deposit, share



INFRASTRUCTURE

Data archives, repositories,
access, licensing, preservation,
cloud, DOI



GOVERNANCE

Funder, University, publisher,
Research institutions, policy,
people, activities, procedure





LIGUE DES BIBLIOTHÈQUES
EUROPÉENNES DE RECHERCHE
ASSOCIATION OF EUROPEAN
RESEARCH LIBRARIES

Ten recommendations for libraries to get started with research data management

Final report of the LIBER working group on E-Science / Research Data Management

10 Recommendations for libraries to get started with research data management

Regrouping the recommendations – Areas of engagement

Support services

#1 Offer **research data management support**, including data management plans for grant applications, intellectual property rights advice and information materials. Assist faculty with data **management plans** and the integration of data management into the **curriculum**.

#6 Support the **lifecycle** for research data by providing **services for storage, discovery and permanent access**.

Infrastructure & standards

#2 Engage in the development of **metadata and data standards** and provide **metadata services** for research data.

#5 **Liase and partner with researchers**, research groups, data archives and data centers to foster an interoperable infrastructure for data access, discovery and data sharing.

#7 Promote **research data citation by applying persistent identifiers** to research data.

#8 Provide an **institutional Data Catalogue or Data Repository**, depending on available infrastructure.

#10 Offer or mediate secure **storage** for dynamic and static research data **in co-operation with institutional IT units** and/or seek exploitation of appropriate cloud services.

Policy & disciplinary practices

#4 Actively participate in **institutional research data policy development**, including resource plans. Encourage and adopt open data policies where appropriate in the research data life cycle.

#7 ... (with some disciplinary views on data citation)

Skills & staffing

#3 Create **Data Librarian posts** and develop **professional staff skills** for data librarianship.

#1 ... and the integration of data management into the **curriculum**.

Areas of RDM engagement

Support
services

Infrastructure
& standards

Policy &
disciplinary
practices

Skills &
staffing

13

Some highlights of LIBER Recommendations

RDM support on data management plans for existing grant, grant applications and advice on intellectual property rights.

- Reference interview management, knowledge of RDM principles
- Knowledge of institutional and extra-institutional resources
- Knowledge of researchers' needs, knowledge of available materials

Provide metadata services for research data

- Metadata skills

Some highlights of LIBER Recommendations



Encourage and adopt open data policies.

Knowledge of researchers' needs, knowledge of available material

Audit interview, knowledge of RDM principles, metadata, licensing



Participate in institutional research data policy developments

Strategic understanding and influencing skills

Knowledge of institution

Some highlights of LIBER Recommendations

Liaise and partner with researchers, research groups, data archives and centers to foster and infrastructure for data access, discovery and sharing.

- Knowledge of institution
- Knowledge of RDM principles, relevant technologies and processes, metadata

Support the lifecycle for research data by providing services for storage, discovery and access.

- Knowledge of RDM principles, relevant technologies and processes, metadata
- Metadata skills

Some highlights of LIBER Recommendations

Provide an institutional Data Catalogue or Data Repository, depending on available infrastructure.

- Audit interviews, knowledge of RDM principles, metadata, licensing
- Knowledge of RDM principles, relevant technologies and processes, metadata

Get involved in subject specific data management practice

- Understanding of RDM best practices as they apply to relevant disciplines, pedagogic skills
- Knowledge of institutions

Support & Consultancy

Researcher (Email; Project title)	Data Stewards (Email address)	Data Stewards (Email address)
AP TPR Dr Goh Hong Ching gohhc@um.edu.my Critical analysis of marine planning model applications	Puan <u>Hanani Fauzi</u> hananif@um.edu.my	Puan Siti <u>Mawarni Salim</u> sitimawarni@um.edu.my
Prof Dr <u>Hazreen bin Abdul Majid</u> hazreen@um.edu.my Project title - TBC	Puan <u>Nuratiqahnadzira Abdul Rani</u> nuratiqah.ar@um.edu.my	Puan <u>Zaharah Ramly</u> zaharahr@um.edu.my
Dr Lee Yean Kee (Project Officer Prof <u>Saadah</u>) yeanke@um.edu.my Project title - TBC	Puan Fairuz <u>Nawwar Mansor</u> fairuznawwar@um.edu.my	Dr Tan Hsiao Wei tanhw@um.edu.my
AP Dr Sabzali Musa Kahn sabzali@um.edu.my <u>Perlaksanaan Program Terapi Seni Visual Sebagai Pemangkin Terhadap Pemantapan Latihan Kemahiran Vokasional Orang Kurang Upaya</u>	Puan Siti <u>Norfateha Azwa</u> norfateha@um.edu.my	Cik Nik Nur <u>Asilah Nik Shamsuddin</u> nnans@um.edu.my
Prof Dr Hanafi <u>Hussin</u> hanafih@um.edu.my RECONCEPTUALIZATION THE WEAVING ART	Cik Aruna J.E <u>Thambidorai</u> aruna@um.edu.my	Puan <u>Sutarmi Kasimun</u> sutarmi@um.edu.my

What is DATA MANAGEMENT PLAN (DMP)



A DMP is a document addressing requirements and practices for managing the project's data, code and documentation, throughout the data life cycle



i.e from the initial planning until the project ends and beyond.



It outlines the data management strategies in a project.



Making plans for how you will collect, document, organize, and preserve your data are all part of the data management strategy.

Support/tools for DMP

- <https://researchdata.um.edu.my/>

EXAMPLE OF DMP

- ..\..\Downloads\DMP-Noorsaadah-THW-ZR-NSJ
(1).docx

DCC Metadata page



Because good research needs good data

About ▾

News ▾

Events ▾

Services ▾

Guidance ▲

Briefing Papers

How-to Guides

Case Studies

Policy Analysis ▾

Metadata ▲

Disciplinary
Metadata

Curation Lifecycle
Model

Data
Management
Plans

Research ▾

Publications ▾

FAQ ▾

Information for ▾

Change cookie settings

Disciplinary Metadata

While data curators, and increasingly researchers, know that good metadata is key for research data access and re-use, figuring out precisely what metadata to capture and how to capture it is a complex task. Fortunately, many academic disciplines have supported initiatives to formalise the metadata specifications the community deems to be required for data re-use. This page provides links to information about these disciplinary metadata standards, including profiles, tools to implement the standards, and use cases of data repositories currently implementing them.

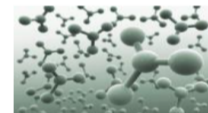
For those disciplines that have not yet settled on a metadata standard, and for those repositories that work with data across disciplines, the General Research Data section links to information about broader metadata standards that have been adapted to suit the needs of research data.

Please note that a [community-maintained version of this directory](#) has been set up under the auspices of the Research Data Alliance.

Search by Discipline



Social Science & Humanities



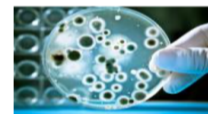
Physical Science



General Research Data



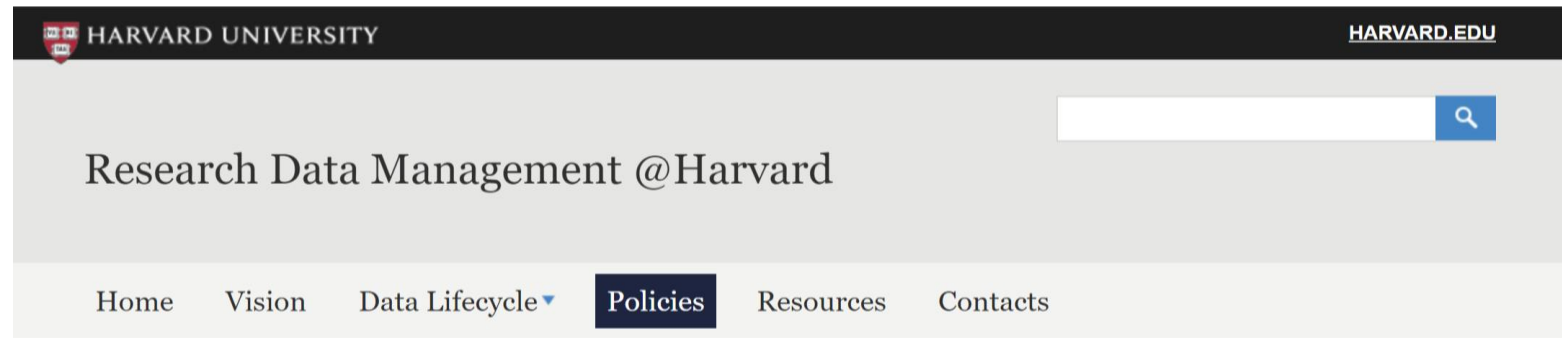
Earth Science



Biology

Policy from Institution

<https://researchdata.management.harvard.edu/policies>



HARVARD UNIVERSITY HARVARD.EDU

Research Data Management @Harvard

Home Vision Data Lifecycle Policies Resources Contacts

HOME /

Data Policies

There are a number of policies and regulations that may impact Harvard researchers working with data. Below is a list of the more commonly applicable internal and external regulations.

Harvard Policies

- [Data Ownership](#): Applies to research data resulting from projects conducted at the University, under the auspices of the University, or with University resources.
- [Data Use Agreements](#): Policy and Guidance documents describe the roles, responsibilities and processes associated with DUAs.
- [Enterprise Information Security](#): University-wide policy applicable to all data created, shared, accessed or otherwise used by Harvard researchers.
- [Genomic Data](#): Policy and procedures for human genomic data sharing and use.
- [Intellectual Property](#): Statement of policy in regard to inventions, patents, and copyrights developed by Harvard researchers.
- [Legal Agreements Workflow and Signature Authority](#): Outlines which offices have authority to review and/or sign specific types of agreements.
- [Open Access](#): Resources pertaining to the schools' policies on Open Access.
- [Publications](#): Overview of acceptable restrictions on publication and review and escalation processes.
- [Research Data Security](#): Applies to all research data physically housed at Harvard or stored remotely under the management of Harvard researchers. [Examples](#) of Research Data Security Levels.
- [Retention of Research Data and Materials](#): Basic principles to guide the retention and maintenance of research records by Harvard researchers and staff.

Federal and State Data Regulations

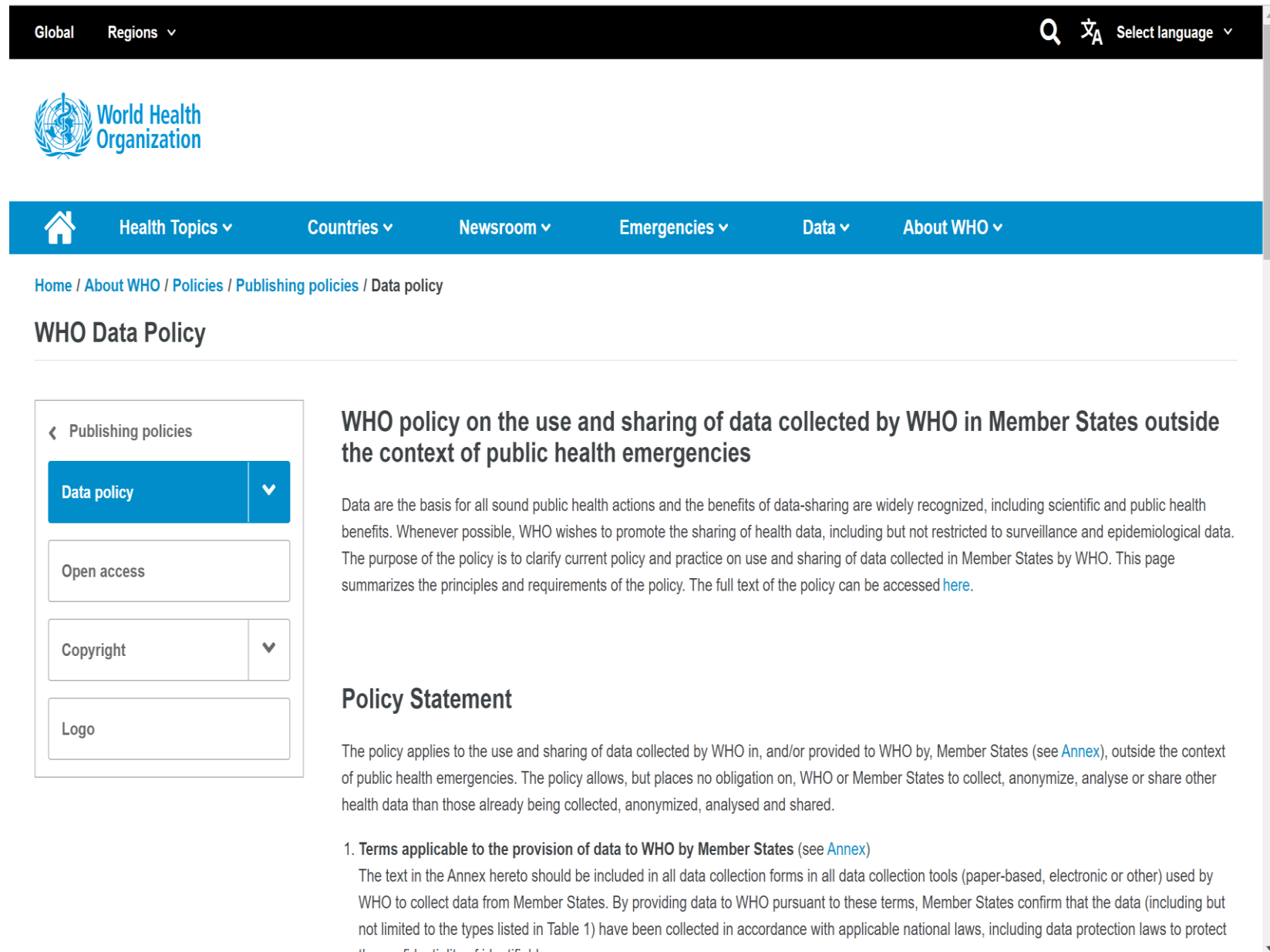
- [Family Educational Rights and Privacy Act \(FERPA\)](#)
- [Federal Information Security Management Act \(FISMA\)](#)

Serving the Nation. Impacting the World.



Policy from Institution/Organization

<https://www.who.int/about/policies/publishing/data-policy>



The screenshot shows the WHO website's 'WHO Data Policy' page. At the top, there is a navigation bar with 'Global' and 'Regions' dropdowns, a search icon, and a 'Select language' dropdown. Below this is the WHO logo and a secondary navigation bar with 'Home', 'Health Topics', 'Countries', 'Newsroom', 'Emergencies', 'Data', and 'About WHO' dropdowns. The breadcrumb trail reads 'Home / About WHO / Policies / Publishing policies / Data policy'. The main heading is 'WHO Data Policy'. On the left, a sidebar menu under 'Publishing policies' includes 'Data policy' (highlighted), 'Open access', 'Copyright', and 'Logo'. The main content area features the title 'WHO policy on the use and sharing of data collected by WHO in Member States outside the context of public health emergencies', followed by an introductory paragraph and a 'Policy Statement' section. The policy statement begins with 'The policy applies to the use and sharing of data collected by WHO in, and/or provided to WHO by, Member States (see Annex), outside the context of public health emergencies. The policy allows, but places no obligation on, WHO or Member States to collect, anonymize, analyse or share other health data than those already being collected, anonymized, analysed and shared.'

WHO policy on the use and sharing of data collected by WHO in Member States outside the context of public health emergencies

Data are the basis for all sound public health actions and the benefits of data-sharing are widely recognized, including scientific and public health benefits. Whenever possible, WHO wishes to promote the sharing of health data, including but not restricted to surveillance and epidemiological data. The purpose of the policy is to clarify current policy and practice on use and sharing of data collected in Member States by WHO. This page summarizes the principles and requirements of the policy. The full text of the policy can be accessed [here](#).

Policy Statement

The policy applies to the use and sharing of data collected by WHO in, and/or provided to WHO by, Member States (see [Annex](#)), outside the context of public health emergencies. The policy allows, but places no obligation on, WHO or Member States to collect, anonymize, analyse or share other health data than those already being collected, anonymized, analysed and shared.

1. Terms applicable to the provision of data to WHO by Member States (see [Annex](#))

The text in the Annex hereto should be included in all data collection forms in all data collection tools (paper-based, electronic or other) used by WHO to collect data from Member States. By providing data to WHO pursuant to these terms, Member States confirm that the data (including but not limited to the types listed in Table 1) have been collected in accordance with applicable national laws, including data protection laws to protect

March 5, 2023

Dataset Open Access

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

 Banda, Juan M.;  Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena;  Chowell, Gerardo

Version 156 of the dataset. FUTURE CHANGES: Due to the imminent paywalling of Twitter's API access this might be the last full update of this dataset. If the API access is not blocked, we will be stopping updates for this dataset with release 160 - a full 3 years after our initial release. It's been a joy seeing all the work that uses this resource and we are glad that so many found it useful.

The dataset files: `full_dataset.tsv.gz` and `full_dataset_clean.tsv.gz` have been split in 1 GB parts using the Linux utility called `Split`. So make sure to join the parts before unzipping. We had to make this change as we had huge issues uploading files larger than 2GB's (hence the delay in the dataset releases). The peer-reviewed publication for this dataset has now been published in *Epidemiologia an MDPI journal*, and can be accessed here: <https://doi.org/10.3390/epidemiologia2030024>. Please cite this when using the dataset.

Due to the relevance of the COVID-19 global pandemic, we are releasing our dataset of tweets acquired from the Twitter Stream related to COVID-19 chatter. Since our first release we have received additional data from our new collaborators, allowing this resource to grow to its current size. Dedicated data gathering started from March 11th yielding over 4 million tweets a day. We have added additional data provided by our new collaborators from January 27th to March 27th, to provide extra longitudinal coverage. Version 10 added ~1.5 million tweets in the Russian language collected between January 1st and May 8th, gracefully provided to us by: Katya Artemova (NRU HSE) and Elena Tutubalina (KFU). From version 12 we have included daily hashtags, mentions and emojis and their frequencies the respective zip files. From version 14 we have included the tweet identifiers and their respective language for the clean version of the dataset. Since version 20 we have included language and place location for all tweets.

The data collected from the stream captures all languages, but the higher prevalence are: English, Spanish, and French. We release all tweets and retweets on the `full_dataset.tsv` file (1,389,125,796 unique tweets), and a cleaned version with no retweets on the `full_dataset-clean.tsv` file (359,764,311 unique tweets). There are several practical reasons for us to leave the retweets, tracing important tweets and their dissemination is one of them. For NLP tasks we provide the top 1000 frequent terms in `frequent_terms.csv`, the top 1000 bigrams in `frequent_bigrams.csv`, and the top 1000 trigrams in `frequent_trigrams.csv`. Some general statistics per day are included for both datasets in the `full_dataset-statistics.tsv` and `full_dataset-clean-statistics.tsv` files. For more statistics and some visualizations visit: <http://www.panacealab.org/covid19/>

More details can be found (and will be updated faster at: https://github.com/thepanacealab/covid19_twitter) and our pre-print about the dataset (<https://arxiv.org/abs/2004.03688>)

As always, the tweets distributed here are only tweet identifiers (with date and time added) due to the terms and conditions of Twitter to re-distribute Twitter data ONLY for research purposes. They need to be hydrated to be used.

This dataset will be updated bi-weekly at least with additional tweets, look at the github repo for these updates.
Release: We have standardized the name of the resource to match our pre-print manuscript and to not have to update it every week.

Preview

emojis.zip

! The previewer is not showing all the files

..
o ..
.. extracted elements

208,630 201,138
views downloads

See more details...

Indexed in

OpenAIRE

Publication date:

March 5, 2023

DOI:

DOI: [10.5281/zenodo.7700410](https://doi.org/10.5281/zenodo.7700410)

Keyword(s):

social media twitter nlp covid-19 covid19

Published in:

Epidemiologia: 2 pp. 315-324 (3).

Related identifiers:

Continued by

<http://www.panacealab.org/covid19/> (Other)

Supplement to

<https://arxiv.org/abs/2004.03688> (Preprint)

Alternate identifiers:

10.3390/epidemiologia2030024 (Journal article)
https://github.com/thepanacealab/covid19_twitter
(Software)

Communities:

BioHackathon
Coronavirus Disease Research Community - COVID-19
Zenodo

License (for files):

Other (Public Domain)

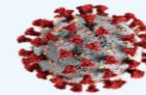
Versions

Version 156 10.5281/zenodo.7700410	Mar 5, 2023
Version 155 10.5281/zenodo.7679153	Feb 26, 2023
Version 154 10.5281/zenodo.7655041	Feb 19, 2023

We expect a 5min planned interruption on the service on March 16th at 06:00 UTC, due to an upgrade of our storage. Thank you for understanding.

Featured communities

Need help uploading? Contact us



Coronavirus Disease Research Community - COVID-19

This community collects research outputs that may be relevant to the Coronavirus Disease (COVID-19) or the SARS-CoV-2. Scientists are encouraged to upload their outcome in this collection to facilitate sharing and discovery of information. Although Open Access articles and datasets are...

Curated by: Covid19_Team, OpenAIRE

Browse New upload

Recent uploads

March 5, 2023 (v156) Dataset Open Access

View

Gene Ontology Data Archive

Carbon, Seth; Mungall, Chris

Archival bundle of GO data release.

Updated on March 9, 2023

52 more version(s) exist for this record

March 5, 2023 (v156) Dataset Open Access

View

A large-scale COVID-19 Twitter chatter dataset for open scientific research - an international collaboration

 Banda, Juan M.;  Tekumalla, Ramya; Wang, Guanyu; Yu, Jingyuan; Liu, Tuo; Ding, Yuning; Artemova, Katya; Tutubalina, Elena;  Chowell, Gerardo

Version 156 of the dataset. FUTURE CHANGES: Due to the imminent paywalling of Twitter's API access this might be the

Need help?

Contact us

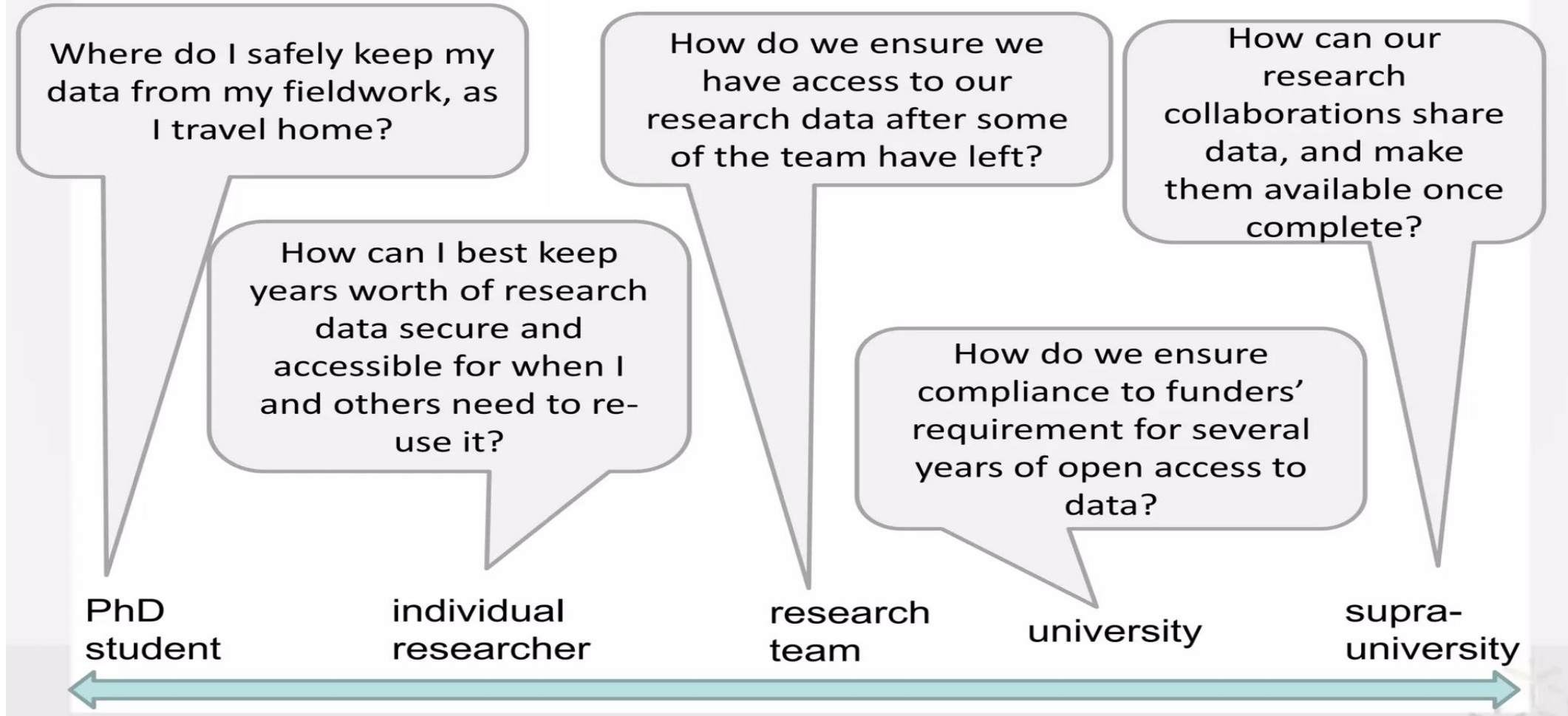
Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:

- Uploading your research data, software, preprints, etc.
- One-on-one with Zenodo supporters.
- Quota increases beyond our default policy.
- Scripts for automated uploading of larger datasets.

Why use Zenodo?

Seeking the real win + win + win + win + win.....



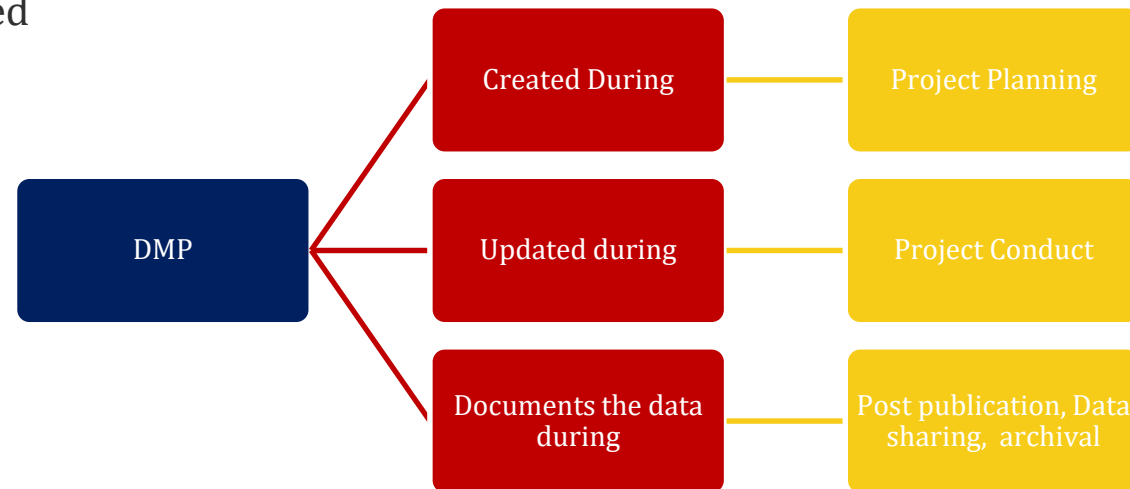
Tony Weir, Director, IT Infrastructure, UoE

Research Data Lifecycle

Identify the data to be collected or used to answer the research question

Plan for data management throughout the lifecycle

This is the stage at which a **Data Management Plan (DMP)** would be created



Many public funders of research ask for a DMP to be submitted as part of a research application

PLAN

COLLECT

PROCESS

ANALYSE

PRESERVE

SHARE

RE-USE



UNIVERSITI
MALAYA

Serving the Nation. Impacting the World.

Research Data Lifecycle

Experiments are carried out, observations made, surveys undertaken, secondary materials acquired

What

- ⑩ Format, Type, Volume

Why

- ⑩ Analyse
- ⑩ Quality Assurance

When

- ⑩ Research Cycle: Create, Process, Analysis & Re-Use

How

- ⑩ Methodologies: Observational, Experimental, Simulation, Derived or compiled, Interview, Survey, Focus Group, Brainstorming, Crowdsourcing
- ⑩ Physical and Digital (folders, files and version control)

Involve documentation of data collection instruments and methods and information necessary to interpret and use the data

PLAN

COLLECT

PROCESS

ANALYSE

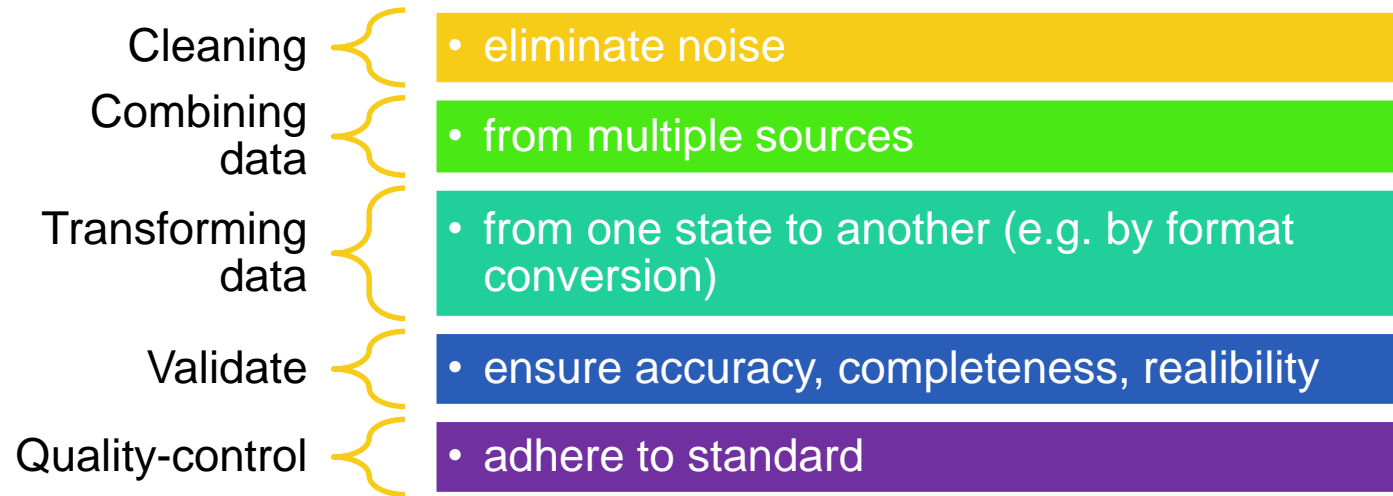
PRESERVE

SHARE

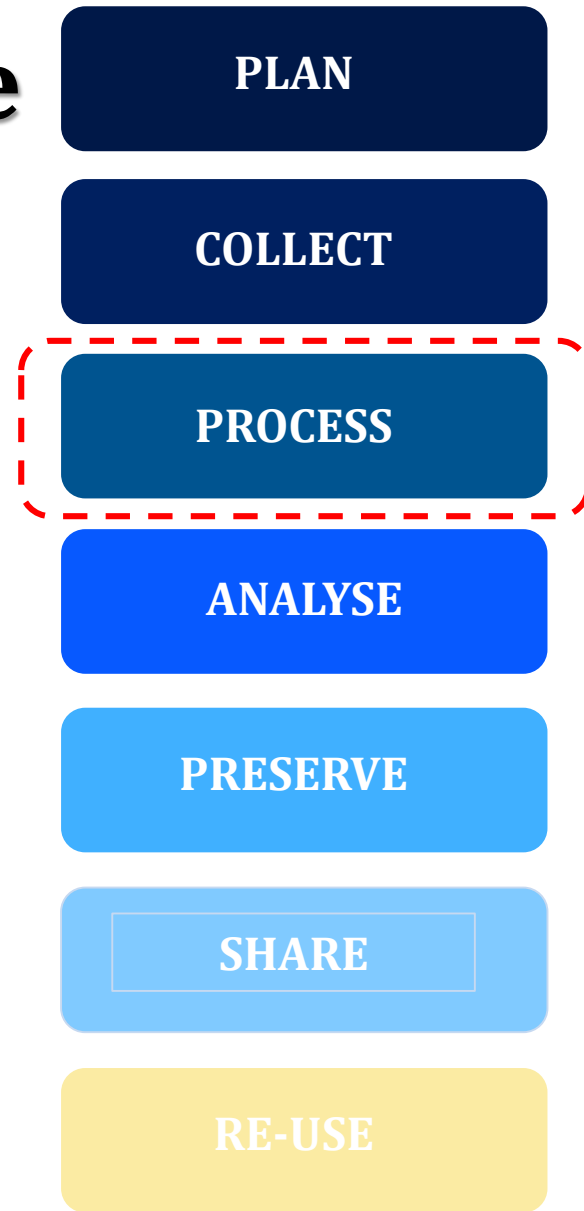
RE-USE

Research Data Lifecycle

Data once collected needs to be processed in order to be usable:



Any data processing needs to be documented for the end result could be replicated from the raw data



Research Data Lifecycle

PLAN

COLLECT

PROCESS

ANALYSE

PRESERVE

SHARE

RE-USE

the raw materials of research are interrogated to produce the insights that constitute the research findings which will be written up and published in research outputs

Instruments and methods used for analysis should be documented

code written for purposes of data analysis and visualisation needs to be preserved and made available in support of research results

Research Data Lifecycle

PLAN

COLLECT

PROCESS

ANALYSE

PRESERVE

SHARE

RE-USE

Maintaining access to data and files over time

WHICH

- Data for short and long term preservation

HOW

- Use policy, legal, institutional, ethical considerations
- Data storage : preserving data in a secure location
- Data Backup: preserving copies of data in separate physical locations

WHAT

- Repository (Institutional or Open), Digital Archive

WHEN

- After completion of budget, digitization, and metadata processes: Data cleansing, verification, validation, archiving

WHY

- Data security from theft or natural disaster, IP security. Easy access by yourself and others

Preservation activities

quality assurance of data

file format conversion

creation of metadata records

assignment of Digital Object Identifiers (DOIs) to datasets

licensing datasets for re-use

access controls (embargo)

Long-term preservation : archiving

Archiving is a necessary first step toward data sharing

Research Data Lifecycle

Data may be made publicly available, or with restrictions where data are of a sensitive or confidential nature

Data held locally or in non-public locations should be managed for others to discover and get access

A data repository will enable discovery of the data

PLATFORMS TO SHARE DATA

Open Data Platform

- ⑩ Software that makes it easier to share and manage open data on the Web
- ⑩ Guide publishers through the procedure of getting data published
- ⑩ Offer consistency and ease of access to open data from around the world to users

Type

- ⑩ Data catalogue
- ⑩ Data management

PLAN

COLLECT

PROCESS

ANALYSE

PRESERVE

SHARE

RE-USE



UNIVERSITI
MALAYA



Why is the reuse of data important?

Why is the reuse of data important?

1

Avoid doing new, unnecessary experiments

2

Run analyses to verify reported findings are correct - making subsequent findings more robust

3

Make research more robust - aggregating results obtained from different methods or samples

4

Gain novel insights by connecting and meta-analysing datasets

Q&A

Serving the Nation. Impacting the World.



www.um.edu.my



[universityofmalaya](https://www.facebook.com/universityofmalaya)



[unimalaya](https://www.instagram.com/unimalaya)



[uniofmalaya](https://www.youtube.com/uniofmalaya)



UNIVERSITI
MALAYA