F1000

# Open data demystified: the essential toolkit for researchers

# Contents



# Introduction

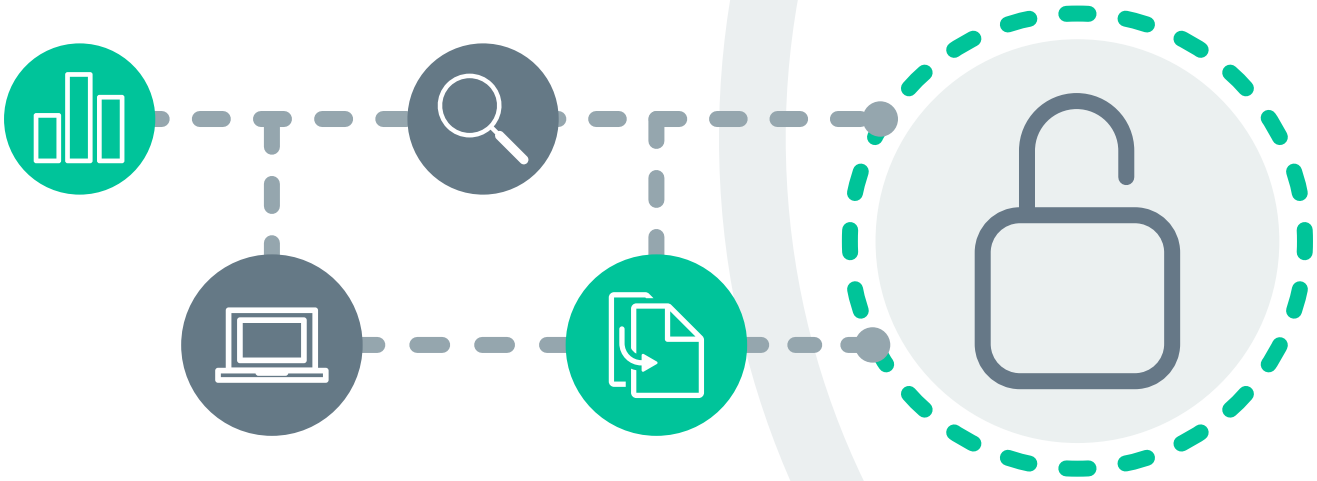
Open data plays a vital role in the research landscape by accelerating the pace of discovery, promoting data reuse, and enabling the testing and validation of research findings. Over the past decade, **data has become a priority for academic stakeholders, including governments, funders, institutions, and publishers worldwide**. Most recently, the White House Office of Science and Technology Policy released a memorandum requiring that researchers make the data the office funds publicly available. But they aren't the first funder to do so. 52 funders listed on Sherpa Juliet require data archiving as a condition of funding, while a further 34 encourage it. Likewise, **today nearly all major scientific journals have an open data policy**. It's clear that data sharing is here to stay, but are researchers ready for the additional time, skills, and effort it requires?

According to the 2022 State of Open Data report, **4/5 respondents favor making research data openly available as standard practice**; but many academics find sharing data difficult. As a result, there is a consensus that training is needed to ensure that the global uptake of open data runs in parallel with the mandates emerging worldwide.

At **F1000,** we recognize that publishers have a huge role to play in supporting researchers to share their data openly. **Open data demystified: the essential toolkit for researchers** is a starting point for any researcher wanting to learn more about data sharing and best practices. In this eBook, you'll find expert guidance on collecting, storing, formatting, sharing, and publishing your data. Together, we can make "data available upon request" a thing of the past and **make data as open as possible but as closed as necessary.**

SECTION 1:

# Making your data open and FAIR



## Identifying your data

The first step to opening your data is to identify all the data you will produce as part of your research project. Research data is the evidence that underpins your findings and can be used to validate the claims made in your publication. There are many types of research data, both quantitative and qualitative. **Common examples of research data vary by discipline but can include:**

- **Survey results**
- **Software**
- **Models and algorithms**
- **Interviews and transcripts**
- **Images, videos, and audio files**
- **Genome sequences**

If you find yourself wondering which parts of your work are research data, The Turing Way suggests asking yourself:

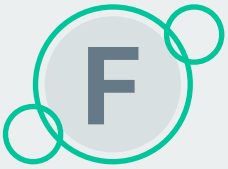
1. **What information do I need to answer my research question?**
2. **What information would I need to back up my findings?**
3. **What information would someone else need to reproduce my work?**

✓ **The answer is your data.**

# Understanding the FAIR principles

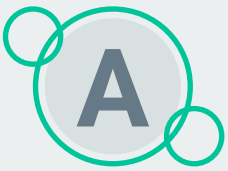
A central purpose of open data is to make research findings reusable so that others can build upon and validate the research. The FAIR principles are a widely endorsed set of guidelines that help to ensure your research data is as reusable as possible. Familiarizing yourself with these principles is essential to improving the **F**indability, **A**ccessibility, **I**nteroperability, and **R**eusability of your data (FAIR).

**Let's take a closer look at the features and behaviors that can contribute to making research data FAIRer:**



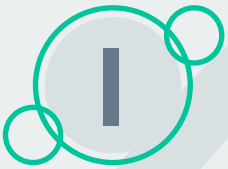

## Findable

Data should be deposited in a repository, giving you a digital object identifier (DOI) or persistent identifier (PID). Use metadata to give a detailed description of your data.




## Accessible

The repository must use a standard protocol like http://. The repository must continue to provide a landing page and metadata even if the dataset is removed.




## Interoperable

The metadata used to describe the data are based on the standard subject vocabularies and should be machine-readable. You can find the subject standards at FAIRsharing.org.


## Reusable

The metadata which describes the data is accurate and relevant. An explicit data license has been applied to the data, explaining what other users can and cannot do.

We know that making your research data FAIR is easier said than done, so **here are some ideas to get you started:**

## 1 Start with a management plan

A data management plan (DMP) is a valuable starting point for collecting or researching data. Creating a detailed DMP before you begin your research, and updating it regularly throughout the research cycle, will help ensure that your research outputs are as open and FAIR as possible at the end of the project. Some funders require grantees to share their DMP as part of their funding application or after the funding has been secured.

## 2 Describe your data clearly

Clearly and accurately describing how you created your data, how it is structured, and what it means is crucial to making your data comply with the FAIR Guidelines. Someone unfamiliar with your data should be able to understand what it is about using only the metadata and documentation provided.

## 3 Preserve your data

Data preservation ensures that your data will be accessible and reusable in the future.
**Best practices for data preservation include:**

- Backing up data files regularly
- Storing master copies of data files in open formats
- Validating preserved data files regularly
- Using more than one form of storage for data files
- Appropriately securing data physically or on any network or computer they are held on

## FAIR data resources

**For further guidance on making your data FAIR, explore the following:**

| | | |
|---|---|---|
| The FAIR principles for the guidelines in full | A FAIRy tale for a simple breakdown of each principle | GO FAIR materials and workshops |

# Why share research data?

Open data are essential for a more equitable society and a level playing field. However, this results in more administrational processes for researchers to open their outputs.

At this point, you might wonder if open data is worth the time and effort it requires. Here's a rundown of how your career, the research community, and society can benefit from data sharing.

## Benefits for your career

### Open data can:

**Increase the discoverability of your research**
Linking your open data and published research outputs can increase readership of your research.

**Increase citations**
Research **shows** that articles with links to datasets shared in repositories generated up up to 25% more citations than articles that did not share data in repositories.

**Enhance the credibility of your work**
When the data supporting your findings is openly available, others can replicate your work to validate your results and conclusions.

**Establish ownership and get credit for your data**
Uploading it to a repository allows you to establish ownership through a persistent identifier so other researchers can cite it.

**Facilitate collaboration and new partnerships**
Researchers in your field and beyond can access and use your data, leading to greater collaboration and new research projects.

## Benefits for the research community

### Open data:

**Supports reproducibility**
Open data enhances research rigor by making it easier for others to validate, replicate, and reproduce your findings.

**Reduces research waste**
When data is openly available, research becomes more efficient by removing duplication of efforts from other researchers.

**Enables others to reuse your data**
Sharing data can lead to reuse by providing a foundation for others to build on.

**Preserves data more securely over time**
Data hosted on a repository is more secure than on a website or personal files.

## Benefits for society

### Open data:

**Gives greater visibility over results of publicly funded research**
Open data offers a chance to make research results openly available as a public good, as research is often publicly funded.

**Can lead to real-world impact**
When data is open, we can accelerate the pace of research discovery to solve societal challenges in real-time.

**Fosters trust in research**
Open data supports transparency and accountability in research, which can help foster public trust in the research process and results.

# Creating a Data Management Plan (DMP)

Hopefully, by this point, you're convinced that sharing your data is the right thing to do and offers profound benefits for your research career. Planning for managing and sharing your data can go a long way in making it easy to open your data at the end of your project.

Before research begins, create a detailed Data Management Plan (DMP). A DMP is a living document that describes how your research outputs will be generated, stored, used, and shared. The document can change and evolve throughout your research project. While most funders and publishers don't require researchers to create a DMP, it can help to ensure efficient data management and makes it easier to make your data FAIR.

According to the DCC, **a thorough Data Management Plan should cover the following:**

## 1 Data collection

**What data will you collect and how?** Here, you can list the file formats you will use to collect, process, and present your data. Consider the following data types when developing your DMP:

- **Raw data:** data collected from the source.
- **Processed data:** a version of the data that has been modified for analysis or visualization.
- **Final data:** data ready to be shared in a publication or repository.

## 2 Documentation and metadata

**What information is needed for the data to be read and interpreted in the future?** Here, you should consider all the information you'll need to describe the data and provide context for your work. Metadata and documentation allow data users to have sufficient information to understand the source, strengths, weaknesses, and analytical limitations of the data to make informed decisions when using it.

Documentation may include data dictionaries, codebooks, protocols, logbooks or lab journals, README files, research logs, analysis syntax, algorithms, and code comments.

## 3 Ethical and legal considerations

**Have you obtained appropriate consent for data sharing?** Here, you'll need to consider whether you have worked with humans, animals, or plants, and if so, what requirements are necessary to ensure you can be ethically and legally compliant when sharing your data.

You should also state who will own the copyright of any data that you will collect or create, along with the license for its use and reuse.

## 4 Storage and backup

**How will you store the data and ensure it is backed up?** Here, you'll have multiple factors to consider, including where the data will be storied; how many copies will be made; who will be responsible for backing up the data; and if your institution provides automatic backup services, to name a few.

If you choose to use a third-party service, you should ensure that this does not conflict with any funder, institutional, departmental or group policies, for example in terms of the legal jurisdiction in which data are held or the protection of sensitive data.

## 5 Data preservation

**How will you preserve the data, and how long?** Decide which data to keep and for how long. This could be based on any obligations to retain certain data, the potential reuse value, what is economically viable to keep, and any additional effort required to prepare the data for data sharing and preservation.

Remember to consider any additional effort required to prepare the data for sharing and preservation, such as changing file formats. Additionally, consider how datasets that have long-term value will be preserved and curated beyond the research lifecycle.

## 6 Data sharing

**How will you share the data?** The methods used to share data will be dependent on a variety of factors including, the type, size, complexity, and sensitivity of data. Here, you should consider these factors and determine the best repository for your research data. There are a wide variety of repositories to choose from, including institutional, general, and discipline-specific repositories.

## 7 Roles and responsibilities

**Who will be responsible for the various data management and sharing tasks throughout the research project?** Map out who will take on all activities, including data capture, metadata production, data quality, storage and backup, data archiving, and data sharing.

You should also carefully consider any resources needed to deliver the plan. This includes, software, hardware, technical expertise, etc. Where dedicated resources are needed, these should be outlined and justified.

## 8 Budget

**Are there any costs associated with sharing your data that you need to prepare for?** Some funder data sharing policies request researchers include the costs associated with data sharing in their DMP. The NIH, for example allows researchers to request funds toward data management and sharing in the budget and budget justification sections of their grant applications.

SECTION 2:

# Storing and organizing your data

**Data loss is a common issue that can harm your research project. Backing up your data frequently and choosing suitable storage solutions can prevent data loss.**

## Where to store your data

### Your institution's local storage solutions

You should always check if your institution provides a network drive you can use to store your data. Such local storage solutions might have policies or guidelines restricting what you can use. So, always familiarize yourself with your local policies and recommendations.

### Portable storage media

Portable storage media, including USB sticks, can pose risks and cause data loss or damage.

### Cloud storage

Cloud storage facilitates storage, backup, and retrieval of data. Before using them for your research data, refer to the terms of use. If, for example, your research involves personal or sensitive data, it is essential to check that the cloud option complies with any data protection rules that apply to the data. For more security, you can also encrypt devices and files where possible.

### Databases and repositories

When you are ready to share your data with the broader community, you can search in FAIRsharing for databases and repositories suitable for your data and access types.

# How to organize your data

## Folders

Creating a folder structure or reusing a previous structure that has worked for you can **help organize your data to support findability**.

It helps if you have enough folders and subfolders to avoid storing files in the wrong folder too much data in a single folder. Plus, your folder structure must be clear and organized following specific criteria, such as data type, analysis methods, who generated the data, when it was generated, and so on.

## File naming conventions

Think of how you should structure your file names and **develop a template**. For example, you could name your files after they were created. This way, you will be able to sort your files chronologically and have a unique identifier per file.

You can use other factors or criteria to name your files, such as the researcher's name, the file type, or the date range of an experiment. Try and keep your names short and avoid any special characters.

Plus, you can explain the file naming convention in a README.txt file so that other researchers who access your data can know what the files are about.

## Backups

A good rule of thumb is to **have 2 or 3 copies of your files stored on at least 2 storage media in different locations.** Ideally, backing up data should be automated and follow your institution's guidelines.

The more often your datasets change, the more you should back them up. If you have a large volume of files and backup proves to be time-consuming, challenging, or expensive, you should set up a list of criteria for when you should back it up. You can include this in your Data Management Plan (DMP).

# Version control

Many research projects involve large files, such as input data or analysis results, which can change over time. **Tracking the subset or version of data a particular analysis or results relies on is essential to support reproducibility.**

Let's say we have a dataset that is the basis for computing a scientific result. If the dataset changes, results will likely become invalid, or scripts based on file names that change between versions will break. Plus, reproducing the original results of the analysis might not be possible if new data replaces the original data without version control in place.

Therefore, it becomes evident that version-controlling data and large files, in general, can support the reproducibility of your research and verify the source of the results. Data identified in precise versions is part of the research outcome and all other research project components.

## Valuable tools for version-controlling data

There are several tools available to manage version controlling and sharing larger files. Some integrate well with repositories and can extend their capabilities to version control large files. With such tools, you can add extensive data to a repository, version control it, revert to previous states, modify it, or even share it as a small-sized file.

**Tools for version-controlling data include:**

- Git LFS
- git-annex
- DataLad

SECTION 3:

# Formatting your data

**An open dataset's 'format' describes how the data is organized and made accessible to humans and machines.** The proper format is essential to ensure data management, accessibility, and reuse.

Here, we will examine how to format two common types of research data: spreadsheet data and sensitive data.

## Spreadsheets

Spreadsheets are used for data entry, organization, analysis, and visualization. By following best practices when using spreadsheets, you help ensure your data is interoperable and reusable for both humans and machines in the future.

## Spreadsheet metadata

A data dictionary is a separate file where each variable is defined, including units and ranges, and often includes other helpful information for interpreting the dataset. A data dictionary supports reuse and reproducibility by helping others (and your future self) better understand your data. A data dictionary should accompany each spreadsheet.

## Do's of spreadsheet data

- Put variables in columns and observations in rows.
- Keep your raw data raw; calculations and analyses should be done in a copy of the file.
- Give each column a descriptive heading that does not include spaces, numbers, or special characters.
- Differentiate between zero and null values.
- Validate your data.
- Keep a separate. TXT file with a title and a legend describing your dataset and outlining any steps you take to tidy your data.
- Use a version control system and back up your files.
- Export each data file in an open non-proprietary format, such as CSV or TAB, with a name that appropriately reflects the content of that file.
- Check your data thoroughly; your data should receive the same care as your publications.

## Don'ts of spreadsheet data

- Put more than one piece of information in a cell.
- Use color coding, embedded charts, comments, or tables.
- Include special (i.e., non-alphanumeric) characters within the spreadsheet, including commas.
- Use merged or blank cells.
- Create multiple worksheets within a spreadsheet.

# Sensitive research data

We recommend authors strive to make their data as open as possible and as closed as necessary. However, we recognize that openly sharing data may not always be feasible due to ethical considerations or third-party restrictions.

Datasets that contain personal data can often be shared by ensuring you have informed consent for data sharing, have applied appropriate anonymization techniques, or controlled access to the data.

# Consent

Research data can be made available for future reuse by ensuring you seek participants' consent. You should inform participants how the research data will be stored, preserved, shared, and reused long-term, and you will maintain confidentiality.

Consent procedures must be tailored for the specific research context, methods, and sample, the nature of the data (personal, sensitive, level of detail), the format of the data (surveys, written, recordings), and the planned data uses and handling. This will influence the type of consent and consent process used. You can find detailed guidance on gaining written or oral consent from the **UK Data Service**.

Researchers must seek consent, but participants can decide if and how their data is shared. Researchers should impartially advise participants about the risks and benefits of research participation and data sharing. Participants then decide what they will consent to, and researchers must honor the participants' wishes. If the participants don't consent, you can't make your data available for reuse.

# Data anonymization

Anonymization alters direct identifiers (name, postal code, phone number, etc.) and indirect identifiers (occupation, gender, location, etc.) so individuals cannot be identified in a dataset. Both quantitative and qualitative data can benefit from anonymization using the following techniques:

# Remove

Remove direct identifiers from your dataset. Where direct identifiers such as a personal name cannot be easily removed while maintaining usefulness of the dataset, use pseudonymization instead (for qualitative data). Better still, plan and avoid collecting identifiable data that is not needed.

## Generalize

Where possible, replace disclosive information with more generalized information – while still maintaining meaning. For instance, a reference to a Catholic Church could be generalized to a place of worship.

## Aggregate and reduce

Variables such as age and location can be aggregated or reduced to decrease the precision of the variable. For instance, recording birth year rather than birthdate is a form of reduction.

## Log

Keep a record of each anonymization step, and keep this file separate from the anonymized data file.

## Reassess

Continually assess whether you have removed the risk of disclosure as you conduct the anonymization process.

## Controlled access

There may be cases where data cannot be fully anonymized, and openly sharing data is not feasible. It may still be possible to make your data accessible, in line with the FAIR Principles, to authenticated users via a controlled-access repository. This depends on what the ethical board approving your study said about data sharing and the level of permission granted. Bear in mind that varied levels of access control can be implemented; you can combine controlled access to potentially disclosive data while openly sharing non-disclosive data.

## Key things to remember when anonymizing your data

- **Be sure to apply an appropriate level of anonymization.** Anonymization may impact the usefulness of data. If data is becoming completely unusable, you might choose to use a controlled access repository instead of continuing to remove things from the dataset.
- **Consider any linked datasets that are (or will be) available.** In some cases, anonymized datasets can be at risk of disclosure when combined with other open datasets. Should this be the case for your dataset, controlled access may be required.
- **Search and replace is useful for anonymizing datasets but be cautious**; search and replace will not apply to misspelled words and may replace words not intended.

SECTION 4:

# Sharing your data

## Open data licenses

Researchers can apply "open" licenses to grant others permission to reuse their research data with minimal restrictions. Permitting reuse supports reproducibility and transparency in research and enables others to build on your findings. Applying an open license to your research data permits others to use your data under copyright law. From the reusers' perspective, the presence of a license answers the question, "What can I do with this dataset?

### What license should you use to publish open access?

**The following open licenses permit maximum reuse by others with minimal restrictions:**

#### Creative Commons Public Domain Dedication (CC0)

CC0 is a public domain dedication tool that is ideal for sharing data. This license allows datasets to be reused for commercial purposes. It has no restrictions on reuse at all, so reusers can distribute, remix, adapt and build upon the dataset in any medium or format with no conditions. While there are no requirements for attribution, citing CC0 datasets is widely accepted and expected in a research context.

#### Creative Commons Attribution Only (CC-BY)

The CC-BY license requires the data to be attributed to the creator. It allows reusers to distribute, remix, adapt and build upon the materials in any medium or format as long as the creator is credited. This license also allows datasets to be reused for commercial purposes. You must make sure that your chosen license adheres to any relevant funder, institutional, legal, or ethical obligations

# Open data repositories

While an open license outlines what others can and cannot do with your data, data repositories provide online storage for researchers to store research data and other research outputs.

Depositing your data in a publicly accessible, recognized repository that assigns a globally persistent identifier ensures that your dataset remains available to humans and machines in a usable future. Funders and journals often maintain a list of endorsed repositories for your use. Still, choosing the best repository from such lists can often be daunting. The type of repository you should use depends on the nature of your research. **Repositories can be categorized into three types:**

## Discipline-specific repositories

Research data differs significantly across disciplines. Discipline-specific repositories offer specialist domain knowledge and curation expertise for particular data types. A discipline-specific repository can make your data visible to others in your research community.

## Institutional repositories

Many institutions offer support to their employees for managing and depositing data. Institutional repositories that accept datasets provide stewardship, helping to ensure that your dataset is preserved and accessible.

## General data repositories

General data repositories accept datasets regardless of discipline or institution. These repositories support various file types and are particularly useful where a discipline-specific repository does not exist.

# Metadata

Metadata is data about data. Complete metadata allows researchers to locate data they created and recall the circumstances and context under which they created and analyzed the data. It also allows researchers outside the original research team to discover, understand and use the data.

Good metadata is associated with the dataset it describes and is available in a machine-readable format. Metadata standards exist to guide how your data and metadata should be structured, formatted, and annotated. These standards can vary by discipline, so choosing the suitable metadata standard for your field of study is essential. The RDA Metadata Standards Directory lists hundreds of standards, extensions, tools, and use cases. You can browse the directory by discipline and subject area.

When depositing your data in a repository, you must fill in as many fields as possible, as this information usually contributes to the metadata record(s). Some repositories require researchers to submit metadata files alongside the data.

## Essential tips for writing metadata

- **Consider the type of research and the nature of the data**, as this can impact what kind of documentation is necessary.
- **Consider the documentation and metadata required**, as it can vary according to the project and the range of people the data needs to be understood.
- **Use recognized community metadata standards** to make it easier for datasets to be combined.

SECTION 5:

# Publishing your data

**When it comes to publishing your research, don't let your data be an afterthought.** Here's how to write a suitable data availability statement and publish a Data Note to increase the potential impact of your work.

## What is a data availability statement?

A data availability statement is a required section of the manuscript that tells the reader how, where, and under what conditions the data associated with your research can be accessed and reused.

## What type of data should it cover?

Your statement should reference all data associated with your article and details of any software you used to process results. Your data availability statement could include up to four sections, depending on the types of data used. **If these apply to your research, you must include them in your statement:**

### Source data

This is data you have not collected yourself as part of your study but has been used for analysis. Source data includes data that you obtained from a third party.

### Extended data

Extended data are additional materials that support the key claims made in your article but are not required to follow the study design and analysis. Examples of extended data include questionnaires and supporting images or tables. You should upload extended data to an online repository.

### Underlying data

Underlying data is data you collected or produced as part of your study. You should upload underlying data to an online repository.

### Reporting guidelines

Where mandatory reporting guidelines apply, you should upload a copy of the relevant guidelines to an online repository.

**At F1000, we recommend uploading all data (except third-party data) to an approved online repository as a single dataset**, which you can then cite throughout your article and in the data availability statement. If you have multiple data deposits, you will need to list all of them as individual datasets.

**For repository-hosted data, your data availability statement must include the following**:

- Repository name
- Title of the dataset
- DOI
- List of all data items (including the full file name and a description of its contents)
- Data license

# Data availability statement examples

Now, let's look at what different types of data availability statements look like in practice.

## Single dataset

**If you have submitted your research to a repository as a single dataset, you can include a straightforward data availability statement like the example below.**

### Data availability

#### Underlying data

Open Science Framework: Japan PVD 2018. https://doi.org/10.17605/OSF.IO/QW2AF; registration DOI https://doi.org/10.17605/OSF.IO/7Y4AV (Yamada, 2020).

This project contains the following underlying data:

- PVDJapan2018.xlsx. (The dataset.)
- Description of Dataset.txt.

Data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

## Multiple data types

**If your research involves different data types, we strongly advise you to use subheadings in your data availability statement for clarity.** This **Research Article** published on F1000Research has source and underlying data from different locations. As a result, the authors have split out their data availability statement using subheadings to clarify this.

### Data availability

#### Source data

Structures of natural compounds were downloaded from the ZINC Database.

Crystal structures of COVID-19 main protease were downloaded from the Protein Data Bank, accession numbers 6LU7 (in complex with N3) and 6Y7M (with OEW).

#### Extended data

Harvard Dataverse: Replication Data for: Computational screening for potential drug candidates against SARS-CoV-2 main protease. https://doi.org/10.7910/DVN/GYFXA0[57].

This project contains the following extended data:

- 2D interaction maps of all OEW pharmacophore-like ligands (PNG).
- 2D interaction maps of all Remdesivir pharmacophore-like ligands (PNG).
- 2D interaction maps of all Hydroxychloroquine pharmacophore-like ligands (PNG).
- 2D interaction maps of all N3 pharmacophore-like ligands (PNG).

Extended data are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

## Third-party data

If your data belongs to a third party, it won't necessarily be stored in an open repository. In this case, **authors should describe precisely how they gained access to the data so that readers can do the same.**

This Research Article uses third-party source data owned by the Demographic and Health Surveys (DHS) Program. Because the authors do not own this data, they have clarified how readers can access the dataset.

### Data availability

#### Source data

The data for this study is owned by the DHS Program. The Individual Recode datasets for the PDHS 2006–07, 2012–13 and 2017–2018 were used for this study and can be obtained here: https://www.dhsprogram.com/data/available-datasets.cfm?ctryid=31

The electronic data is available from the DHS Program under its terms of use. Before downloading the data, users must register as a DHS user for reasons laid out on the DHS Program website and dataset access is only granted for legitimate research purposes.

## Sensitive data

Suppose your research involves sensitive data which can't be shared openly due to ethical, data protection, or confidentiality considerations. In that case, you should still **include a data availability statement and clearly state why you could not share the data**.

This F1000Research Data Note describes a dataset containing sensitive details of intensive care patients, which you cannot upload to an open repository for ethical reasons. They have explained this in their Data Availability Statement and included instructions on how readers can apply for access to this dataset.

### Data availability

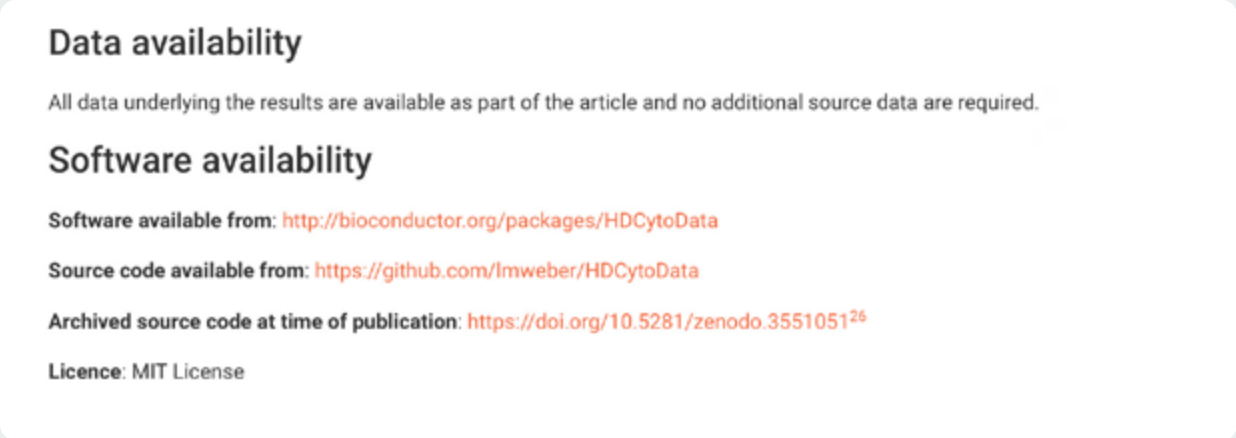
#### Underlying data

The sensitive nature of these data means that they are only available internally to UHB staff for the purposes of clinical audit and service evaluation activities via the CAG guidelines. For external researchers, ethical approval may be obtained via formal application to the NHS Integrated Research Application System (IRAS) for a specific research project. The IRAS website (www.myresearchproject.org.uk) has full instructions; however, interested parties are advised to contact the corresponding author (christopher.bourdeaux@uhbristol.nhs.uk) to discuss the application.

## What about software?

**If describing new software, you should make the source code available on a Version Control System (VCS) such as GitHub or BitBucket and provide details of which open source license you applied.** If you choose to publish with F1000, we will also need a permanent, archived version of the source code at the time of publication to be uploaded to an approved repository.

This F1000Research **Software Tool Article** includes a software availability statement so readers can easily find and access the software described in this article.



Data availability

All data underlying the results are available as part of the article and no additional source data are required.

Software availability

**Software available from**: http://bioconductor.org/packages/HDCytoData

**Source code available from**: https://github.com/lmweber/HDCytoData

**Archived source code at time of publication**: https://doi.org/10.5281/zenodo.3551051[26]

**Licence**: MIT License

## What if my article doesn't have any data?

If no data is associated with your article, you should still **include a data availability statement that makes this clear to readers.** In this case, the statement should read: "No data are associated with this article."

Suppose you present all associated data within your article. In that case, the statement should read: "All data underlying the results are available as part of the article, and no additional source data are required."

# Data Notes

Effective data sharing extends beyond simply depositing your dataset in a repository. **At F1000, we believe everyone should share data in a discoverable, usable, and reproducible way.** That's why we created Data Notes.

**Data Notes are short, peer-reviewed articles that describe how and why you created a dataset. They include:**

- Dataset rationale, protocol, and validation details
- Information about any limitations of the dataset
- Information on where and how to access the dataset as part of a data availability statement
- Reference to the dataset using a formal data citation

Data Notes don't include any analyses or conclusions but promote research data's discoverability and potential reuse by providing a detailed description. Data Notes credit data producers with a citable, peer-reviewed publication and support new research collaborations across disciplines.

LEARN HOW TO PUBLISH A DATA NOTE

# Conclusion

**At F1000**, we recognize that researchers are under increasing pressure to share their data from academic stakeholders, including the government, funders, institutions, and publishers worldwide. In this eBook, we've provided expert guidance on collecting, storing, formatting, sharing, and publishing your data.

If you're ready to start your open research publishing journey, we offer **unparalleled Open Data Support** for submitting authors through our self-service **Resources for Researchers hub**, author **webinars**, and support from our open data specialists. Any author requiring additional support can expect hands-on guidance from our Editorial team to meet our progressive open data policy.

*Journals and funding agencies are now catching up in terms of progressive data sharing policies. Policy changes on data sharing coming from the NIH and OSTP last year will have a huge impact on data sharing in the US, and we're already noticing a trend towards stronger data policies from funders globally. These changes support an open research ecosystem where researchers expect to be asked to share data as they interact with various stakeholders, helping to make data sharing the norm involved too.*

**Rebecca Grant**
Head of Data & Software Publishing, F1000