# A practical guide to data management and sharing for biomedical laboratory researchers

Fouad K*[1], Vavrek R[1], Surles-Zeigler MC[2], Huie JR[3,4], Radabaugh H[3], Gurkoff GG[5,6,7], Visser U[8], Grethe JS[2], Martone ME[2,4], Ferguson AR[3,4], Gensel JC*[9], Torres-Espin A*[1,3]

*Corresponding author

Karim Fouad
karim.fouad@ualberta.ca

John C Gensel
gensel.1@uky.edu

Abel Torres-Espin
espin@ualberta.ca
abel.torresespin@ucsf.edu

[1]Department of Physical Therapy, Faculty of Rehabilitation Medicine, University of Alberta, Edmonton, AB, Canada.

[2]Department of Neuroscience, University of California, San Diego, La Jolla, CA, United States

[3]Department of Neurosurgery, Brain and Spinal Injury Center, Weill Institutes for Neurosciences, University of California, San Francisco, San Francisco, CA, United States

[4]San Francisco Veterans Affairs Healthcare System, San Francisco, CA, United States

[5]Center for Neuroscience, University of California Davis, Davis, CA, United States

[6]Department of Neurological Surgery, University of California Davis, Davis, CA, United States

[7]Northern California Veterans Affairs Healthcare System, Martinez, CA, United States

[8]Department of Computer Science, University of Miami, Coral Gables, FL, United States

[9]Spinal Cord and Brain Injury Research Center and Department of Physiology, University of Kentucky College of Medicine, Lexington, KY, United States

## Abstract

Effective data management and sharing have become increasingly crucial in biomedical research; however, many laboratory researchers lack the necessary tools and knowledge to address this challenge. This article provides an introductory guide into data management, and the importance of FAIR (Findable, Accessible, Interoperable, and Reusable) data-sharing principles for laboratory researchers. We explore the advantages of implementing organized data management strategies and introduce key concepts such as data standards, data documentation, and the distinction between machine and human-readable data formats. Furthermore, we offer practical guidance for creating a data management plan and establishing efficient data workflows within the laboratory setting, suitable for labs of all sizes. This includes an examination of requirements analysis, the development of a data dictionary for routine data elements, the implementation of unique subject identifiers, and the formulation of standard operating procedures (SOPs) for seamless data flow. To aid researchers in implementing these practices, we present a simple organizational system as an illustrative example, which can be tailored to suit individual needs and research requirements.

By presenting a user-friendly approach, this article serves as an introduction to the field of data management and offers a practical guide to help researchers effortlessly meet the common data management and sharing mandates rapidly becoming prevalent in biomedical research.

## Contents

# 1. Background

Data management and sharing is a fundamental part of academic research. We routinely share data with our lab members and colleagues. With new funder and journal mandates, we are now required to share with the broader scientific community. Recognizing the limited reproducibility and replicability in biomedical sciences, all stakeholders have made efforts to introduce solutions to improve transparency in methods and analysis, increase reproducibility, and reduce waste (Bandrowski and Martone, 2016; Begley and Ioannidis, 2015; Chan et al., 2014; Collins and Tabak, 2014; Levesque, 2017). Many biomedical journals now require datasets underlying the paper's claims be submitted with a manuscript or released in data-sharing repositories. Research communities are developing data-sharing initiatives (Callahan et al., 2017; Chervitz et al., 2011; Chou et al., 2022; “Data sharing is the future,” 2023; Fouad et al., 2019; Karpen et al., 2021; Markiewicz et al., 2023; Ohmann et al., 2017; Torres-Espín et al., 2021), and data sharing has become a scholarly field in its own right, with exponential growth in publications containing the terms "data sharing" (Fig. 1). In parallel, funding agencies have implemented policies and recommendations for the sharing of data from the research they support. In addition, they are establishing programmatic strategic plans that push data sharing to fuel the widespread adoption of data-driven technologies such as artificial intelligence (AI) and machine learning (ML). A big challenge for the laboratory researcher is that data management and sharing are not always part of their structured training, and it requires time to navigate and learn the tools. Data sharing often feels like an inconvenience without a clear initial payoff, as it has become a complex ecosystem to navigate, with constantly changing scenarios, policies, and rules. However, data management within the laboratory accrues immediate benefits to the laboratory itself and greatly facilitates ultimate outside sharing. This document offers a guide to the laboratory researcher, introducing concepts and providing a step-by-step example of how research data management best practices can be embraced, from experimental design to data collection and sharing. This will assist in complying with the increasing requirements of good data stewardship and make data management and sharing a part of the laboratory research endeavor rather than an afterthought (Dempsey et al., 2022; Martone and Nakamura, 2022).

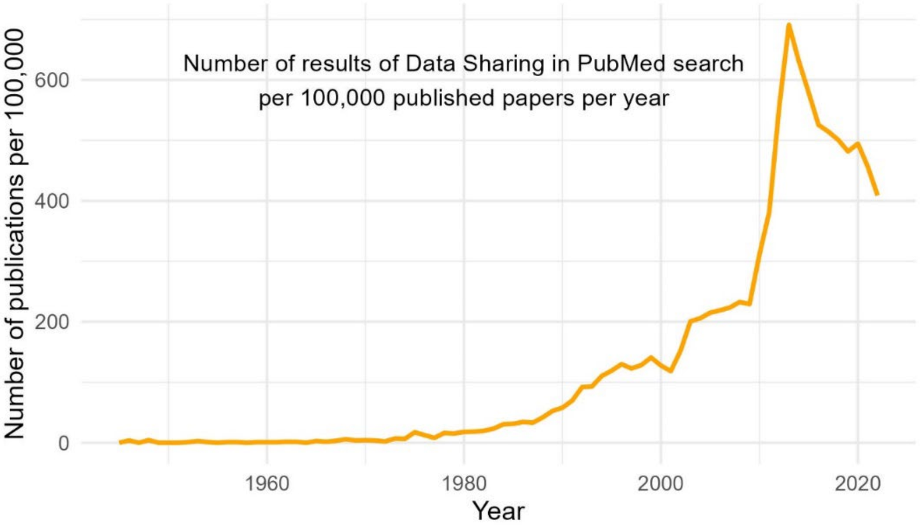


**Figure 1.** The number of publications when searching "Data Sharing" in PubMed has exponentially grown in the last three decades compared to the growth of total number of published papers in PubMed. Results from 1945 to 2022 shown as ratio of number of results in PubMed search per 100,000 published papers. Data obtained using PubMed by Year (Sperr, 2016).

*Data sharing requirements and mandates*

Governments and Foundations supporting research are increasingly aware of the value of research data, the importance of transparent and open publishing, data reuse, and the need for policies to foster data stewardship best practices, and to reduce waste. This is in response to an international movement promoted by the intergovernmental Organisation for Economic Co-operation and Development (OECD), dedicated to promoting economic progress and world trade. In 2006, the OECD adopted the "Recommendation of the Council concerning Access to Research Data from Public Funding" (OECD, 2006), signed by the governments of 41 countries. The OECD states:

*"The Recommendation seeks to assist governments, research support and funding organisations, research institutions, and researchers in dealing with the barriers to and challenges in improving the international sharing of research-relevant digital objects…"*

In 2021, these recommendations were updated to incorporate further details on the relevant digital objects:

*"... the reuse of data is increasingly and critically dependent on the availability of related metadata, as well as bespoke algorithms, workflows, models, and software (including code), which are essential for their interpretation. Providing access to these digital objects, in addition to the data itself, is essential..."*

Consequently, funding agencies have started recommendations and policies for data management and research data sharing. For example, the US National Institutes of Health (NIH) has begun mandating that most researchers and institutions receiving public funds must make their data publicly available through the issue of its Policy for Data Management and Sharing (NIH DMS policy, 2023). The NIH DMS policy "*emphasizes the importance of good data management practices and establishes the expectation for maximizing the appropriate sharing of scientific data generated from NIH-funded or conducted research, with justified limitations or exceptions*." It encourages prospective planning for data management and sharing in human and non-human research by requiring DMS plans at the time of grant submission. Similarly, the Canadian Tri-Agency, composed of The Canadian Institutes of Health Research (CIHR), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Social Sciences and Humanities Research Council of Canada (SSHRC), is in the process of incremental implementation of their policies on Digital Data Management (Government of Canada, 2021). These include requiring data management plans in selected grant applications, mandating research institutions to establish a research data management strategy, and a general demand for data sharing.

The list of policies and mandates is extensive and depends on funding and agency jurisdiction. The Sherpa Juliet project from Jisc (formerly known as Joint Information Systems Committee), a UK not-for-profit organization for digital services and solutions for high education and research, provides a repository of open access policies by countries and institutions (Sherpa Juliet, 2023). Beyond governmental funding agencies, foundations and other funders are also starting to recommend or require data sharing. Thus, reading the data policies related to relevant funding agencies is highly recommended. The following sections provide descriptions and a guide on the key elements to understand the policies and help with their implementation.

## 1.1.Improving shared data: the FAIR data principles

In 2014, at a workshop at the University of Leiden, the acronym FAIR (Findable, Accessible, Interoperable, and Reusable) was coined, and a set of data principles was developed. These principles guide the process of sharing data, making them findable, accessible, interoperable, and reusable by **both humans and machines** (Wilkinson et al., 2016). Since then, the term FAIR and its principles have been endorsed and recommended by journals, scientific communities, and funding agencies, and they are becoming the go-to guiding principles for developing data-sharing strategies.

The FAIR data guidelines were explicitly designed to facilitate and enhance the reusability of research data. In short, data that is shared needs to be findable and accessible, meaning that data cannot be stored/hidden in "file drawers", known as "dark data", as it prevents its reuse beyond the data creators (Ferguson et al., 2014; Scargle, 1999; Schembera and Durán, 2020). Once you find data and can download or otherwise access it, you should be able to reuse it. This means data needs to meet a minimal set of standards and provide enough information to make them more understandable by more than just the data creators. For effective reuse of data, it must be interoperable, meaning data has an expected structure facilitating the ability of machines to use the data while supporting its reuse by humans. FAIR specifies that data should be consumable by *both* humans and machines. With the exponential increase in the amount of collected research data and the size of datasets, the concepts of machine-readable and reusability are critical considerations.

Let's take the example of 'omics. With the advent of high-throughput omics methods, the scientific community quickly realized that having data standards and well-annotated formats for files containing data were essential to allow for big-data transfers, pre-processing, and reuse by different software and to provide a common language for all researchers to understand shared data (Chervitz et al., 2011). Nowadays, across all data types, the volume of data generated is several times higher than just a few years back, and the complexity of studies has also increased (Fire and Guestrin, 2019). Therefore, ad hoc and traditional data organization procedures in the laboratory may not be sufficient for successful management. Therefore, the need for efficient machine-readable data and sharing standards can be expanded to any biomedical domain, from neuroimaging to electrophysiology to clinical data, to name a few (Hicks et al., 2013; Markiewicz et al., 2023; Maumet et al., 2016; Niso et al., 2018; Rübel et al., 2016).

# 2. The advantages of organized data collection, managing, and sharing

Collecting and managing well-organized and structured data ensures efficiency and productivity along the research data lifecycle (e.g., planning a study, collect and process data, analyze, archive and share, and reusing data; Fig. 2). It enables researchers to comply with funding mandates, journal requirements, work with collaborators, and the future use of data by the data creators themselves and others; generally, any activity that requires accessing, sharing, and using data (Dempsey et al., 2022). Many have experienced the desire to explore new ideas, knowing that the data to do so exists somewhere in a filing cabinet (more likely hard drives nowadays), but are unable to find or understand it, especially years after experiments were executed. Imagine having your historical data ready to generate preliminary results for new grant applications. Another example of enhanced efficiency is the need to prepare data to be released with a journal publication, where one might find that much work is needed to get the data ready for compliance. Moreover, what if you could seamlessly identify similar studies, access their data and integrate with your own? Thinking about data management and sharing from the beginning and during experimental design can serve the data creators and the entire research community.

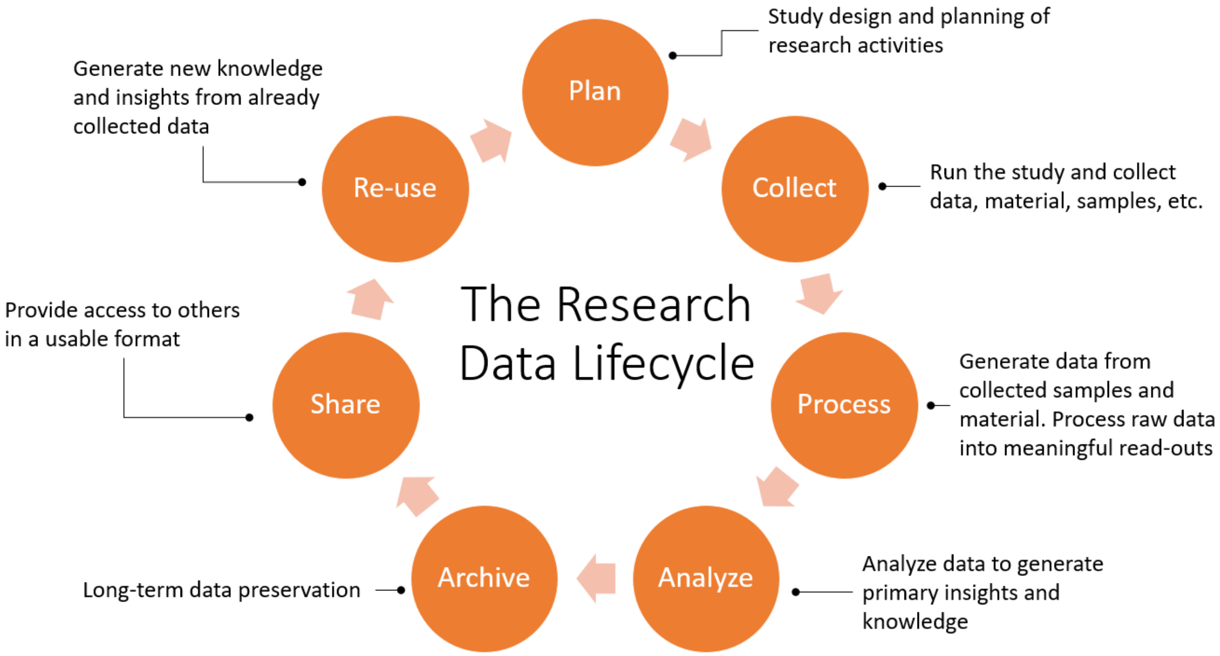

Study design and planning of research activities
Plan
Generate new knowledge and insights from already collected data
Re-use
Collect
Run the study and collect data, material, samples, etc.
The Research Data Lifecycle
Provide access to others in a usable format
Share
Process
Generate data from collected samples and material. Process raw data into meaningful read-outs
Long-term data preservation
Archive
Analyze
Analyze data to generate primary insights and knowledge

**Figure 2.** The Research data lifecycle

## 2.1.Value for the data creator

Well-organized collection, management, and sharing of data have tremendous value, not only for research globally but for those that created the data. Embracing data management and sharing from the initial experimental design and data collection provides easy mechanisms for compliance with the increasing pressure of releasing data. It provides effortless DMS plans for funding applications, it can be used as a standard for training new lab members (e.g., for organizing data or outcome measures defined in the laboratory protocol), and ease reviewing old data and integrating new research. Indeed, in our own experience, laboratory managers and principal investigators have repeatedly returned to the cloud-based data repositories where they deposited data as well-organized archiving systems for their own research. Why? Because they can reliably find, access, and reuse it. Consider the gain in efficiency. Conforming to single procedures on data organization and structure in the laboratory from the beginning reduces time and potential errors during analysis and re-formatting data for publication and sharing. In addition, it increases reproducibility in the lab and between groups. This dramatically facilitates data recovery and use in the data creator's own group, even after personnel changes. However, the value of sharing data goes well beyond administrative purposes. It has been shown that sharing data fosters collaborations and increases citations (Bierer et al., 2017; Colavizza et al., 2020; Kennedy, 2012; Lee et al., 2016), which suggests that data sharing provides another venue for scientific recognition (Bierer et al., 2017; Gorgolewski et al., 2013). By sharing data, the data creators establish themself as transparent and rigorous researchers, which is appreciated in their research communities.

## 2.2.Value for the community

Data sharing serves as a mechanism to increase transparency (Boué et al., 2018; Fecher et al., 2015) and to reduce publication bias towards positive outcomes (Scargle, 1999), as a tool for increasing reproducibility, replicability and rigor (e.g., reducing type I and II inferential errors), generating new knowledge, fostering innovation, and reducing waste and animal numbers in research (Carr and Littler, 2015; Chan et al., 2014; Flanagin et al., 2022; Ioannidis, 2014; Roundtable on Environmental Health Sciences et al., 2016). Having access to data can facilitate knowing what has been done before, thus, avoiding unnecessary replications of studies. It also enhances our ability to pool studies and perform meta-analysis more efficiently and rigorously (Burke et al., 2017; Riley et al., 2010). Shared data also serves as a valuable educational tool. Trainees can learn about data analyses, quantitative literacy, and statistical methods using publicly available datasets. The use of imperfect data and reflection of actual experiments provides trainees with greater insight and adaptability to correctly interpret their own results in the future.

The value of data sharing for the scientific communities can be seen in those sharing data for years. Let's continue with the example of the -omics field, where standards for datasets such as microarray and RNA sequencing experiments were established and broadly adopted by the community. It was rapidly recognized that generating big datasets requires the community's help to fully exploit them with data sharing at its core. The development of new knowledge, data pipelines, and analytical tools, such as the BioConductor project (Gentleman et al., 2004; Hu et al., 2021), has skyrocketed due to the open access to data and software. This experience could be expanded to other fields and types of data, where relatively small datasets are created daily, known as the "long tail of small data" (Ferguson et al., 2014). Widespread structured data sharing enables exploration into topics outside the original experimental goals and permits the creation of new aggregate datasets across multiple sources (Almeida et al., Submitted; Curran and Hussong, 2009; Dhruva et al., 2020; Ferguson et al., 2013; Nielson et al., 2015; van der Steen et al., 2008) Or the use of historical data to better plan new experiments (Hu et al., 2022).

# 3.Concepts and definitions

This section provides an overview of common concepts and definitions researchers initiating good data practices will encounter. The section is divided into data, standards, and documentation. These topics are intertwined and complement each other to provide a complete understanding of a dataset. Table 1 summarizes these concepts.

**Table 1**. Summary of key general concepts and definitions

| **Key General Concept** | **Definition** |
|---|---|
| **Data** (definition extracted from NIH policy in data management and sharing) (NIH DMS policy, 2023) | The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. |
| **Data Standards** | Set of rules on different dataset characteristics such as format, structure, metadata, and definitions that have been agreed upon by a group of people. |
| **Machine-readable vs. Human-readable** | Machine-readable data refers to structured data in a format easily readable by a computer or device rather than humans. How humans organize things may not be the most appropriate for machines to be able to read and make sense of them. Organizing data for digital storage and sharing in machine-readable forms increases interoperability and reusability, reducing waste and inefficiencies. |
| **Data formatting standards** | Set of standards that specify how data should be formatted and structured for a given application, usually to facilitate machine readability and automatic processing. |
| **Data definition standards** | Set of standards that specify what data variables (elements) mean, how they relate to each other, and how to collect them, such that data from different laboratories and studies can be considered equivalent. |
| **Data documentation** | All annexed information about datasets facilitates human and machine understanding. These are key for human readability. |
| **Data dictionary** | File containing the definitions for all variables and measurements, their units, permitted values, and other information at the variable level. |
| **Metadata** | Accompanying information and documentation that provides details about your dataset. The data dictionary can be considered part of the metadata and other information such as abstract, author list, associated methodology, and funding source. |
| **Protocol** | The document outlining the steps-by-step methodology or procedure where all information about a study is structured and well-articulated. |
| **Data management and sharing plan** | The document specifies how data management and sharing will be performed for a study. These are usually required by funding agencies and regulatory bodies, with the same overall goal but unique requirements. |

## 3.1.Data

There is no single definition of what constitutes data. The Oxford Dictionary defines it as *"facts or information, especially when examined and used to find out things or to make decisions*," which might be too broad to be meaningful in practice. The NIH Policy for Data Management and Sharing defines scientific data as:

> *"The recorded factual material commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data do not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects, such as laboratory specimens."*

Based on this definition, each scientific community may establish concrete recommendations for the information needed to validate and replicate research findings. Two critical remarks are the need for sufficient quality data and that data is not restricted to the factual material used to support a publication. Limiting the data definition to be sufficient to validate and replicate findings might narrow the utility and promises of shared data. Leonelli (Leonelli, 2015) provided an overview from the philosophy of science perspective and generated its interpretation:

> *"Data thus consist of a specific way of expressing and presenting information, which is produced and/or incorporated in research practices so as to be available as a source of evidence, and whose behavior and scientific significance depend on the context in which it is used"*

Leonelli suggests that what constitutes data depends on the context of use, where researchers decide what can be used as evidence. For example, in neuroimaging research, the raw sequence MRI files may constitute essential data, while in fields using MRI as a read-out of interest, the outcome of processing those images might be sufficient and more important. This view agrees with NIH's definition above that what constitutes data is a communal acceptance by a set of researchers or a community. Borgman (Borgman, 2015) suggests that data becomes data only when used to support evidence. Adapting this definition, Martone, Garcia-Castro, and VandenBos (Martone et al., 2018) noted data "*... as the measurements, observations or facts taken or assembled for analysis as part of a study and upon which the results and conclusions of the study are based* ". As a broad definition for this guide, data can be considered as the smallest unit of quantities on which evidence is based and generally excludes things like tissue samples, western blot gels, or other physical objects from which the quantities have been obtained. For example, in behavioral neuroscience, videos of animals performing a task might not be data to be shared, but the quantifications of animal performance organized in a digital file used to support scientific claims can be seen as the smallest unit of measure.

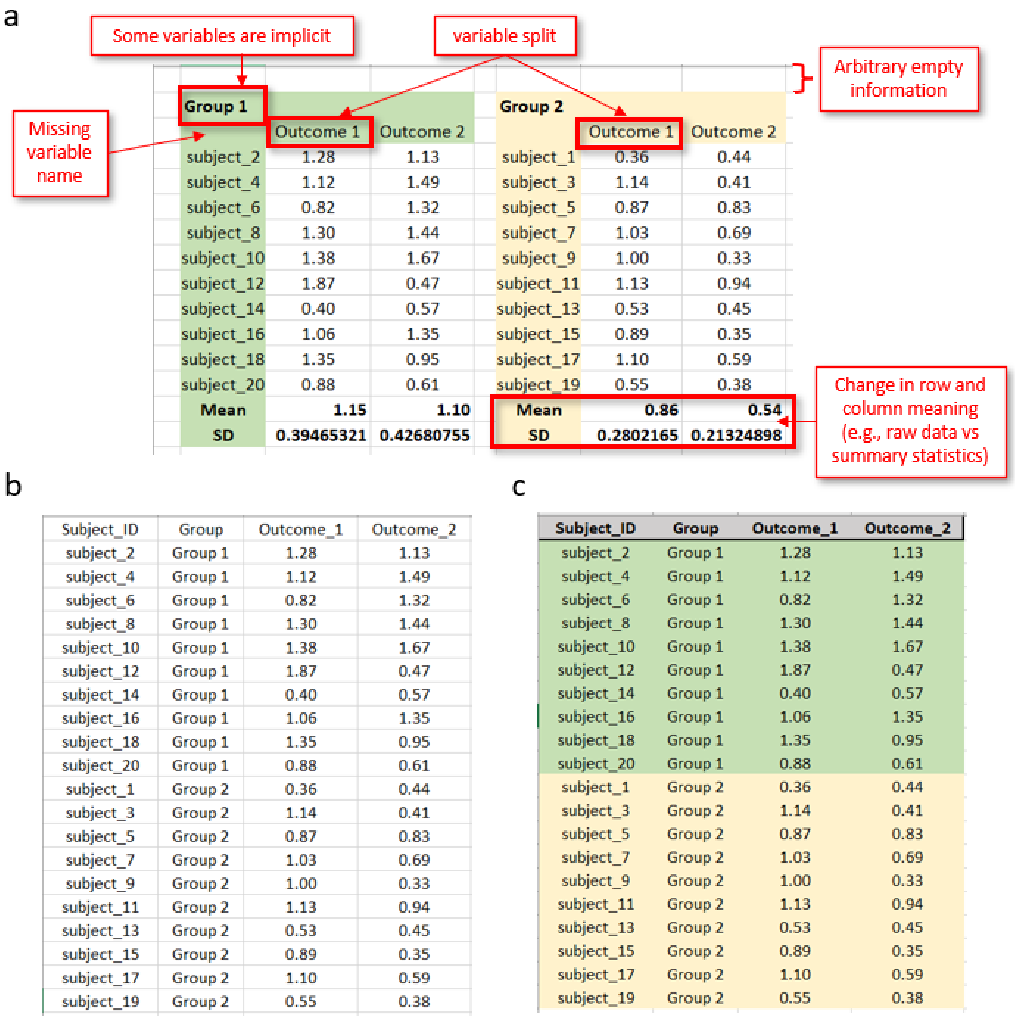
We suggest looking for data standardization, sharing, and validation initiatives in your field of research to have a more concrete idea of what constitutes data to be preserved, managed and shared.

## 3.2.Data standards

Data standards are rules on various dataset characteristics, such as format, structure, metadata, and definitions (e.g., data elements), that have usually been agreed upon by a group of people. Following established standards will help make data more interoperable and reusable and is one of the core practices recommended by FAIR. It also helps to organize data and associated information and establish a systematic process. Comparable to protocols for conducting experiments in a reproducible manner, adopting data standards helps to devise protocols for consistent data collection, storage, and sharing. Several data standards exist depending on the biomedical field, the data type (e.g., genomics, imaging), and how data is being collected. Searching through registries is a good place to start, such as the FAIRsharing.org (https://fairsharing.org/) and the International Neuroinformatics Coordinating Facility (INCF) Standards and Best Practices Portfolio (https://www.incf.org/resources/sbps). Here we define three key concepts related to standards that data practitioners may encounter during the data life cycle.

### 3.2.1.Machine-readable vs. human-readable

There is a difference in how humans and machines read and interpret information. Computers do very well with highly structured and formatted data, particularly when it is always presented in the same way. Humans are more flexible in their approach and can make sense of data even if there is less structure. Part of that flexibility is our ability to understand data's meaning. Although that gives us an advantage when it comes to understanding, that flexibility creates dispersed situations where humans store data in unpredictable or inefficient ways for efficient automation by a computer. Figure 3 shows an example of organizing data in a way that humans understand but that can be challenging to use by machines unless some expected structure is used. This creates a trade-off between data workflow efficiency and human understanding. The term "human- and machine-readable" refers to formatting data in a helpful way for both parties, balancing the trade-off (Wilkinson et al., 2016). This is important since, ultimately, humans are gaining the knowledge captured in data. Discussing all the potential ways to accomplish that balance is beyond the scope of this manuscript. However, it would generally entail predictably formatting data for a computer to operate on while the information is retrievable in a way humans can comprehend. This can be achieved by adopting data standards, accompanying the data with amply documentation and metadata, and developing tools to navigate the trade-off.


a
Some variables are implicit
variable split
Arbitrary empty information
Missing variable name
Change in row and column meaning (e.g., raw data vs summary statistics)
b
c

| Group 1 | | | Group 2 | | |
|---|---|---|---|---|---|
| | Outcome 1 | Outcome 2 | | Outcome 1 | Outcome 2 |
| subject_2 | 1.28 | 1.13 | subject_1 | 0.36 | 0.44 |
| subject_4 | 1.12 | 1.49 | subject_3 | 1.14 | 0.41 |
| subject_6 | 0.82 | 1.32 | subject_5 | 0.87 | 0.83 |
| subject_8 | 1.30 | 1.44 | subject_7 | 1.03 | 0.69 |
| subject_10 | 1.38 | 1.67 | subject_9 | 1.00 | 0.33 |
| subject_12 | 1.87 | 0.47 | subject_11 | 1.13 | 0.94 |
| subject_14 | 0.40 | 0.57 | subject_13 | 0.53 | 0.45 |
| subject_16 | 1.06 | 1.35 | subject_15 | 0.89 | 0.35 |
| subject_18 | 1.35 | 0.95 | subject_17 | 1.10 | 0.59 |
| subject_20 | 0.88 | 0.61 | subject_19 | 0.55 | 0.38 |
| Mean | 1.15 | 1.10 | Mean | 0.86 | 0.54 |
| SD | 0.39465321 | 0.42680755 | SD | 0.2802165 | 0.21324898 |

| Subject_ID | Group | Outcome_1 | Outcome_2 |
|---|---|---|---|
| subject_2 | Group 1 | 1.28 | 1.13 |
| subject_4 | Group 1 | 1.12 | 1.49 |
| subject_6 | Group 1 | 0.82 | 1.32 |
| subject_8 | Group 1 | 1.30 | 1.44 |
| subject_10 | Group 1 | 1.38 | 1.67 |
| subject_12 | Group 1 | 1.87 | 0.47 |
| subject_14 | Group 1 | 0.40 | 0.57 |
| subject_16 | Group 1 | 1.06 | 1.35 |
| subject_18 | Group 1 | 1.35 | 0.95 |
| subject_20 | Group 1 | 0.88 | 0.61 |
| subject_1 | Group 2 | 0.36 | 0.44 |
| subject_3 | Group 2 | 1.14 | 0.41 |
| subject_5 | Group 2 | 0.87 | 0.83 |
| subject_7 | Group 2 | 1.03 | 0.69 |
| subject_9 | Group 2 | 1.00 | 0.33 |
| subject_11 | Group 2 | 1.13 | 0.94 |
| subject_13 | Group 2 | 0.53 | 0.45 |
| subject_15 | Group 2 | 0.89 | 0.35 |
| subject_17 | Group 2 | 1.10 | 0.59 |
| subject_19 | Group 2 | 0.55 | 0.38 |

| Subject_ID | Group | Outcome_1 | Outcome_2 |
|---|---|---|---|
| subject_2 | Group 1 | 1.28 | 1.13 |
| subject_4 | Group 1 | 1.12 | 1.49 |
| subject_6 | Group 1 | 0.82 | 1.32 |
| subject_8 | Group 1 | 1.30 | 1.44 |
| subject_10 | Group 1 | 1.38 | 1.67 |
| subject_12 | Group 1 | 1.87 | 0.47 |
| subject_14 | Group 1 | 0.40 | 0.57 |
| subject_16 | Group 1 | 1.06 | 1.35 |
| subject_18 | Group 1 | 1.35 | 0.95 |
| subject_20 | Group 1 | 0.88 | 0.61 |
| subject_1 | Group 2 | 0.36 | 0.44 |
| subject_3 | Group 2 | 1.14 | 0.41 |
| subject_5 | Group 2 | 0.87 | 0.83 |
| subject_7 | Group 2 | 1.03 | 0.69 |
| subject_9 | Group 2 | 1.00 | 0.33 |
| subject_11 | Group 2 | 1.13 | 0.94 |
| subject_13 | Group 2 | 0.53 | 0.45 |
| subject_15 | Group 2 | 0.89 | 0.35 |
| subject_17 | Group 2 | 1.10 | 0.59 |
| subject_19 | Group 2 | 0.55 | 0.38 |

**Figure 3**. Example of the difference in organizing tabular data on a spreadsheet for human-readability, machine-readability, and a combination of both. Machine-readability is complicated by common strategies to increase human-readability in spreadsheets (**a**). This might include implicit coding of variables using colors, not providing explicit names for variables, splitting information and adding visual cues (e.g., empty columns), and conflicting the meaning of rows and columns (e.g., mixing descriptive statistics with raw data). Although this form of organizing data in a file can be machine-readable, all these strategies are hard to standardize and predict. The same data can be organized in a more structured and repeatable format to facilitate machine-readability (**b**). Markup strategies can also be used to navigate the machine vs. human trade-off (**c**).

### 3.2.2.Data formatting and specifications

One of the most critical aspects of reusing data efficiently is that it is structured and stored similarly. Digital data storage is determined by the type of computer file (file format), their relationship, and how data in those files are organized (formatting). When discussing data formatting, one may come across terms such as "data model" or "data schema." These contain explicit information on how to structure data. Readers familiar with clinical research might have come across "common data models" or CDMs that standardize the formatting and structure of clinical data. The Observational Health Data Sciences and Informatics Observational Medical Outcomes Partnership (OHDSI-OMOP) and the Clinical Data Interchange Standards Consortium Operational Data Model (CDISC-ODM) (Huser et al., 2015) are examples of CDMs. The use of CDMs has shown to be of high value for combining clinical data across different sources, which reduces the cost and time of pooling and analyzing data. Other advantages of using CDMs beyond pooling data are the standardization of processes like quality checks, the transformation from one format to another, automated analysis, etc. A computer can be programmed to perform many different data management tasks if data is consistently formatted. Open formats (i.e., non-proprietary) ensure that a variety of tools can use the data and will be readable even if the software that created it is no longer available. Unfortunately, generalized cross-laboratory data standards are still less common for pre-clinical than for clinical research.

### 3.2.3.What data format standard or CDM do I adopt for my data?

We recommend that you first consider open data formatting standards in your specific field of research, mandated by regulatory bodies or the data-sharing platform you plan to use. For instance, perhaps, researching which NIH-supported repositories might best suit your data if NIH funds you (https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data). Researching such standards would help with the decision and ultimately save the time needed for re-formatting. For example, in neurotrauma research, the Open Data Commons for Spinal Cord and Traumatic Brain Injuries (ODCs; https://odc-sci.org and https://odc-tbi.org) (Callahan et al., 2017; Chou et al., 2022; Fouad et al., 2019; Torres-Espín et al., 2021) require a minimal set of formatting standards; therefore, adoption of these by neurotrauma researchers during data collection and management can significantly reduce the time and friction to upload, share and publish data to the ODCs.

### 3.2.4.Data definitions and standards common reference model

Another crucial aspect of making your data interoperable and reusable is using standard definitions for the same things, such that your data is comparable to the data collected by others. For instance, what one researcher defines as "injury severity" is the same across the entire research community. Agreeing on the definitions is essential. If two datasets contain a variable named "latency", but they mean two different things, this creates a conflict of what a potential aggregated dataset reflects. This is highly challenging in practice because there is generally no single way to define what we do in the laboratory. A solution can be common reference models as standards for defining data variables (a.k.a. data elements). These can include vocabulary and terminologies, ontologies, and common data elements (CDEs). They provide information on naming variables, their definitions, the concepts these variables should relate to, the relationship between concepts, and sometimes define how data needs to be collected or measured to fulfill those definitions (Sheehan et al., 2016). For examples, there are clinical CDEs for TBI (Hicks et al., 2013; Manley et al., 2010; Thompson et al., 2015; Whyte et al., 2010) and SCI (Biering-Sørensen et al., 2015), and the TBI field has been developing CDEs for pre-clinical research

(LaPlaca et al., 2021; Smith et al., 2015). The US National Library of Medicine has a repository of all NIH-endorsed CDEs (https://cde.nlm.nih.gov/home). The BioPortal project has the most comprehensive repository on biomedical ontologies (Whetzel et al., 2011).

## 3.3.Data Documentation

The full value of data can only be realized if there is sufficient information about the processes that were applied to create the data. Therefore, data documentation accompanying any dataset is essential to data management and sharing. Common data documentation includes metadata, data dictionaries, and protocols.

### 3.3.1.Metadata

The definition of what constitutes metadata can vary across research fields. The basic definition is "data about the data" or information that does not constitute the data itself but provides an understanding of different aspects of the data, facilitating its reuse. The most basic metadata types are file size, format, and dynamic range, e.g., 8-bit. However, descriptive metadata goes well beyond the attributes of the data file. For instance, an image file taken with a digital camera may contain associated information about the settings of the camera and the time the image was taken, or keywords associated with a dataset or the date a dataset was uploaded to a repository can be considered part of the metadata. There are different types of metadata, all with different goals. A data dictionary, as described below, can be considered descriptive metadata that provides definitions and other elements for the content of a dataset. The citation of a dataset (similar to the citation of a paper) provides referencing metadata, and a data reuse license may provide legal metadata. Data repositories would generally indicate what information beyond a dataset is required for data uploading and archiving.

An important piece of metadata are persistent identifiers. These are unique references to documents, files, and any digital object that persist in time. Two common ones that biomedical researchers are used to are the Digital Object Identifier (DOI) and the Open Researcher and Contributor ID (ORCID). In our digital era, persistent identifiers are key to make objects findable and accessible over time through internet, reducing the chances of the so called “link rot” or the fact that web links stop working because the address or location changes or disappears.

### 3.3.2.Data Dictionaries

For data to be understood and reused by others (including your future self), users must know what variables were measured and what these measurements represent. A data dictionary (a.k.a. codebook) provides this information in a standard format. Even if you are not planning on sharing your data, it is encouraged and good data management practice to have data dictionaries for your datasets. You may now know what a variable name means in your spreadsheet, but will anyone know when you leave the lab? Will you know if you try to reuse the data two years from now? A data dictionary is a critical lab asset that ensures the data that have taken great effort and resources to acquire will not go to waste in the future due to poor documentation. In addition to providing a significant benefit to you and your lab, a data dictionary is often required by data-sharing repositories and data-sharing mandates promoting interoperability and reusability.

A data dictionary should include the minimum information required to understand a data field or element in the dataset. Depending on the nature of the data element, this may need more or less information. In general terms, a data dictionary should at least include the following:

- **Variable name:** The unique name of a variable or data element in the dataset. Variable names are often short names with abbreviations and other contractions. To facilitate readiness, the variable name is sometimes accompanied by a "title" or "label" entry specifying a long or spelled name. For example, the variable name "subjectID" can be accompanied by the title "study subject identifier." Constructing variable names that are both human- and machine-readable is worthwhile. Special characters such as "%", commas, and spaces in variable names may limit machine readability (Broman and Woo, 2018).
- **Definition or description:** A human-readable narrative explaining the variable and its meaning, how the variable was collected, etc.
- **Units of measurement:** When applicable, the units of measurement are important information. For example, it is essential to know if a variable defining a subject's age is measured in days, months, or years.
- **Permitted values:** The values that the variable or data element can take. For instance, a categorical variable such as sex may take values of "female" and "male." It is often the case that the values of some variables are codified, for example, "female" = 1, "male" = 0. In those cases, adding an entry to the data dictionary specifying the codification is very useful for data interpretation.

Each data repository may have different requirements for a data dictionary, and some repositories may not even ask for one. As stated above, we highly recommend building one accompanying each dataset, as it makes the data use process less cumbersome and helps to standardize outcome measures within a lab and the research community. An example of a specific data dictionary format for odc-sci.org and odc-tbi.org can be seen in the *example of a simple data organization system* section below and supplementary material.

### 3.3.3.Protocols

Biomedical researchers are well-versed in the importance of protocols. In general terms, we can distinguish between a protocol outlining the steps-by-steps of a methodology or procedure and a study protocol where all information about a study is structured and well-articulated before the study starts. Both provide valuable detailed information often not present in other documentation that can increase the understanding of data collection and the nuances of how data has been produced. For example, a manuscript's material and methods section is often a reduced version of a detailed protocol due to word limits and human readability, which may omit important considerations for full reproducibility and understanding of methodologies. These would include the data collection procedures and how the raw data might have been pre-processed or analyzed to produce derived data elements presented in a dataset. Protocols are also an excellent tool for total transparency in the data lifecycle (Fig. 2).

Protocols can also be standardized and shared in association with datasets. Some data repositories will accept extra documentation that can be used to provide protocols. Some other data repositories will suggest depositing protocols in a dedicated database for protocol sharing, such as protocols.io (Teytelman et al., 2016; https://www.protocols.io/), and link them to the data. Journals also allow the citation of externally hosted protocols in the methods section. Recommendations on what to include in a protocol have been previously described (Cameli et al., 2018). As in the case of data dictionaries, setting protocols at the beginning of the data collection can be beneficial for sharing data and maintaining reproducible methods and laboratory institutional memory.

### 3.3.4.Computer Code

It is becoming increasingly common to collect, process, and analyze data by programming computers to take some, if not all, steps in the workflow. The code provides a form of documentation but also a way to be transparent and be able to reproduce the steps. When the code is made in-house (i.e., custom programs by the lab), managing and sharing code jointly with your data is highly recommended and often required, licenses permitted. When proprietary software is used, it is important to document program and operating system versions to reproduce their output. Documenting the shared code, usually by adding non-coding lines explaining what the code does, the inputs, and the expected outputs, is important. In addition, if full code pipelines are shared, well-organized documentation such as a manual of operations becomes essential for its reusability.

There are several ways to share code. Nowadays, it is common to develop and maintain code using version control systems and platforms like GitHub. The finalized version of the code can then be frozen and shared with archiving tools like Zenodo, which creates a persistent identifier (https://docs.github.com/en/repositories/archiving-a-github-repository/referencing-and-citing-content). An advantage of doing it this way compared to, for instance, sharing code as supplementary material to a publication or on personal websites is that if new versions are generated, the reader can more easily access it or be pointed to a specific version of the code. Data repositories will often have recommendations about sharing documentation, including code.

### 3.3.5.Data management and sharing (DMS) plan

Funding agencies often require data management and sharing plans as part of their documentation for grant applications. These vary in format across agencies and countries, but they have in common a formal document explaining how data will be managed during research, archived, and shared. Although data management plans are not new, they are increasingly becoming an essential part of any project application with the recent addition of sharing. Maintaining good data organization and procedures for data management in the lab can help to provide material for grant applications and ease compliance with funding agencies. Thinking through in detail about what data will be generated, e.g., number of data types, size, formats and standards, helps to ensure that you request adequate resources for management, analysis, and sharing.

An example of a data management and sharing plan is the one NIH requires for grant applications. NIH has provided material on the expected content of DMS plans and how to write them (NIH DMS policy, 2023). Table 2 summarizes the elements to include in the NIH DMS plan, but we refer the reader to the NIH resources to write their plans.

**Table 2**. Summary of the different elements in the NIH DMS plan

| DMS plan element | Description |
|---|---|
| **Data type** | A brief description of the data to be managed and shared. This includes a summary of what type of data and modality (e.g., images, genomics), the level of data processing, the amount of data to be shared, and the metadata listing. |
| **Related Tools, Software and/or Code** | An explanation on whether tools and software are needed to access and use shared data and reproduce findings. This should include a description of how to access these tools. |
| **Standards** | A description of the data standards to be used. |
| **Data preservation, access, and associated timelines** | Provide names of the data repositories that will be used. Explain how data is made findable, whether it uses specific identifiers (e.g., digital object identifier or DOI), and the timeline for sharing data. |
| **Access, Distribution, or Reuse Considerations** | Explanation of any factor affecting shared data's access, distribution and/or reuse. For example, the need to apply for qualified access or any regulations that restrict data access and reuse. |
| **Oversight of Data Management and Sharing** | Explanation of how compliance with the DMS plan will be monitored, how often, and by whom at the funded institution. |

### 3.3.6.Data reuse and licensing

The sharing license is one crucial documentation associated with shared data and associated documentation as stated in the FAIR principles. These contain the shared data's legal rules and establish what is and is not allowed with the data. It is important to check under what license and conditions data are being shared since the license type would determine how permissive the sharing is and the legal requirements of potential data reuse. For instance, a Creative Commons 4.0 by attribution license (CC-BY 4.0, https://creativecommons.org/licenses/by/4.0/) allows anyone to copy and distribute the data in any medium or format (share), and they can remix and build upon the data for any purpose, even commercially (adapt). At the same time, someone using the data must give appropriate credit (attribution, citation) and provide a link to the license. However, if the more restrictive CC-BY-NC-ND is used, the data can be only used for non-commercial (NC) purposes, it cannot be modified and distributed to others (no derivatives; ND), and attribution must be provided (BY). Some repositories require a specific license, while others let you choose among various licenses. Suppose your chosen repository allows a choice of license. In that case, discussing the license type with your colleagues before the data are generated is a good idea to avoid misunderstanding.

# 4. Developing a data workflow for your lab

1. **Requirements analysis:** Before embarking on the design of data workflow for an experiment, spending some time compiling information that may exist in your field should be considered. For instance, in the TBI research field, some CDEs have been developed to allow standardizing data collection, and adopting already defined CDEs would greatly facilitate downstream processes. Get familiar with the data repositories common in your field and any existing data standards and vocabularies. The National Library of Medicine maintains a list of repositories where you can share your data. A good source can be checking for the requirements and endorsements by the usual funding agencies and journals for your research.

   Consider the infrastructure that is available to you. Your institution may have a dedicated repository for managing your data or deals with cloud storage services like Google Drive or Box. Having a centralized place where your labs data is managed avoids having data hosted on personal devices like laptops or under personal accounts that may not be accessible after a person leaves.

2. **Creating a standard data dictionary for routine data elements:** Start with the minimum standards from your field if they exist. Making sure you build upon those from the conception and design of experiments will avoid missing required information that may be difficult to collect retrospectively. In addition, it is a good idea to add at least all data dictionary attributes essential for those requirements. Consider that a lab data dictionary will be regularly updated and adjusted as new experiments in a lab require the collection of new types of data. Although it is an initial investment to organize the methods of a laboratory in such a document, it will pay off greatly in the long run. It can be used to train staff and students to ensure data collection is standardized within the lab and over generations of students. Creating templates

for data collection and entry can greatly facilitate adherence by all lab members. Secondly, keeping provenance and version control will help to interpret data, and lastly, a common data dictionary can be easily adjusted to fit the requirements for new data publications. See the example of a data dictionary below.

3. **Anticipating the data format for where it will be shared:** Every repository may have different formatting requirements, and establishing data collection in anticipation of data-sharing formats can save much time. For example, the odc-sci.org and odc-tbi.org require data to be uploaded as .csv (comma-separated values) files, where rows are unique observations and columns are variables/outcome measures or features about those observations. Another important thing to remember when establishing a data collection and management format is the distinction between having data formatted for analysis or data for storage and sharing. It is more convenient to collect and store data in a way that allows for easy transformation to different requirements for analysis, depending on the analysis type and software used. Otherwise, storing data formatted for a specific analysis will make it very difficult to re-format the data for sharing among members of your lab and the community or for conducting further analysis. Similarly, balancing machine vs. human-readable data formatting can greatly facilitate downstream data workflows in your lab.

   Make sure that any system you use helps to manage different versions of the data and that these versions are easily retrievable. For example, saving a named version under the versioning menu in Google Drive makes it easy to go back to that version. Alternatively, you can implement a practice where you designate a "version of record" and then make all subsequent modifications on a copy of that file.

4. **Generate a system for the unique identification of subjects and encourage single-subject data tracking at the start of an experiment:** Research subjects are the most common unit of observation in biomedicine; we collect data from specific subjects. Developing a unique identification system greatly helps keep track of data collection, management, and analysis at the individual subject level. For instance, identifying experimental animals using something like "animal 1" or "animal 2" in each experiment can be problematic since "animal 1" would identify more than one subject across experiments. Those data should not be mixed. This is particularly important for long-term data management and sharing, as the original data collectors move on with their careers, it becomes more difficult to trace the origin of the data. Unique identifiers at a lab, research group, or center level are then part of the good data practices. We give one example of how to generate a unique identifier in the next section.

5. **Create documentation and standard operating procedures (SOPs) for data workflow, including data management and sharing.** Once the data workflow in a lab has been established, we recommend generating documentation and SOPs for every step, from data collection and storage to management and sharing. These can be part of other SOPs in a lab, such as laboratory protocols, that serve as instructions, training material for newcomers, and documentation for grant applications (DMS plan).

# 5. Example of a simple data organization system

This section provides an example data organization system that the authors use, and it can be easily implemented in any laboratory with minimal effort and become part of the SOPs for the lab. We also provide a slide deck as a quick reference (Torres-Espín, 2023; https://doi.org/10.5281/zenodo.8071997). This workflow can be adopted as it is, adapted to specific needs, or combined with more sophisticated solutions such as electronic lab notebooks or full data management systems such as DataLad (Halchenko et al., 2021). This organization system has been designed with data sharing in mind.

## 5.1. Digital data storage and file organization (Fig. 4)

We recommend centralized digital data storage in the cloud and/or locally. Cloud services can be convenient as everyone in the lab can access them, and most institutions may offer at least one option. It is crucial to consider backups of data and private access. A simple but effective system can be used for file organization, as in Figure 3. The organizational unit is the experiment and the subjects in each experiment.

- **Lab documentation.** A folder containing all documents needed for the lab to function (e.g., SOPs, training material, product catalogs)
- **Data.** The main folder is where all the data is stored. Inside this folder, we can find the experimental catalog and sub-folders for each experiment.
  - **Experimental catalog (Fig. 5).** A file to track each experiment and provide basic information. See more below.
  - **Experiment subfolder.** Each experiment has its folder named with a unique identifier (e.g., AA1), and it contains at least five key elements:
- **Summary log.** A document that logs the activities that happen for each experiment.
- **Links and resources:** A log of external sites that may house experiment-specific information, e.g., a GitHub file, and electronic lab notebook record.
- **Subject catalog** (Fig. 6)**.** A document listing and cataloging each subject. See more below.

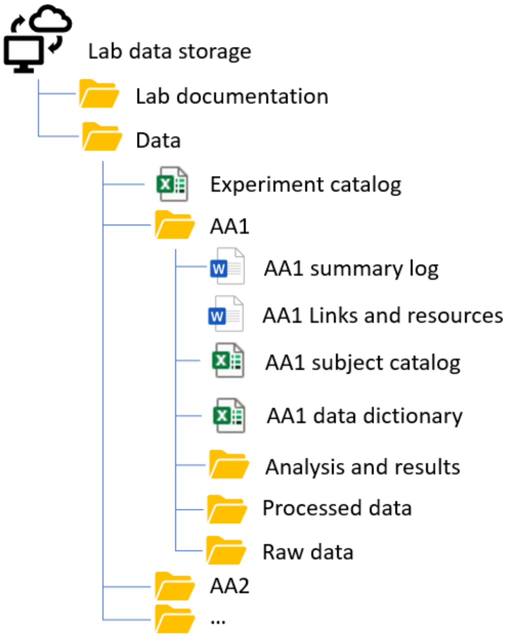


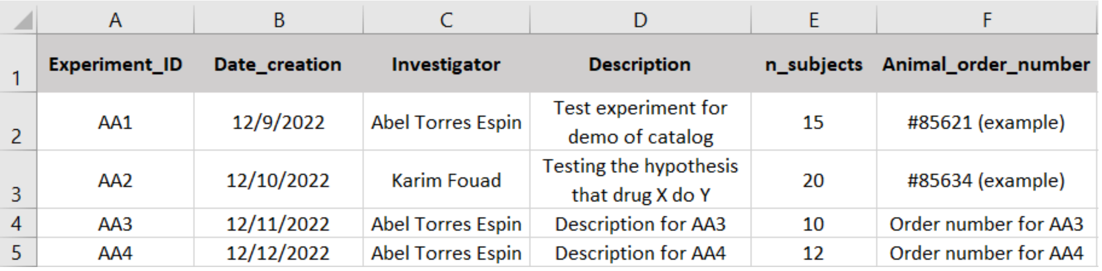
**Figure 4.** Folder organization tree of digital files.

- **Data dictionary** (Fig. 7)**.** Data dictionary for the for experiment AA1
- **Analysis and results subfolder.** A folder containing all the analysis documents (e.g., R scripts, SPSS, SAS, STATA, GraphPad files) and results and outputs of analysis (e.g., graphs, slide presentations).
- **Processed data.** Files containing the processed data. We recommend having files in a sharable format (e.g., csv) and having data dictionaries.
- **Raw data.** This can be the place to store digital raw data such as microscope images, videos, electrophysiology recordings, etc.

## 5.2.The experiment catalog (Fig. 5).

The experiment catalog is crucial in maintaining experiment organization and understanding each subfolder's content. Ideally, as new experiments are planned, these are logged in the catalog, providing provenance. In our example, we build the catalog using a spreadsheet. We suggest keeping track of at least the following basic information for each experiment:

- **A unique experiment identifier (Experiment_ID).** Each experiment has an identifier that serves as a link between the catalog and the respective experiment folder. We used the format *LetterLetterNumber*, AA1, for the first experiment. This allows for an easy increase as new experiments are designed (e.g., AA2, AA3, …, AA9, AB1). Combining two letters and one digit gives a total of 2925 possible experiments. If more are needed, more letters/digits can be used. We recommend a systematic identifier instead of researchers' names, initials, or specific dates, as those are harder to track and can bring confusion. However, we also recommend saving that information so that it is easy to remember these experiments (see investigator and running title below).
- **Date of creation.** The date that the experiment was created in the catalog. This allows for temporal tracking of each experiment.
- **Investigator.** The responsible investigator for the experiment. This is the lead person in charge of the execution and progress of the given experiment.
- **Collaborators:** Names of other lab members and colleagues involved in the experiment.
- **A running title.** This will help to find/identify specific experiments.
- **Description.** A brief description of the experiment and the goals. More details on the experiment are provided in the specific subfolder.
- **The number of subjects.** The number of subjects used for each experiment.
- **Animal order number.** If the experiment requires animals, the animal order number or animal series identification can help link with orders and track animal usage in the lab.



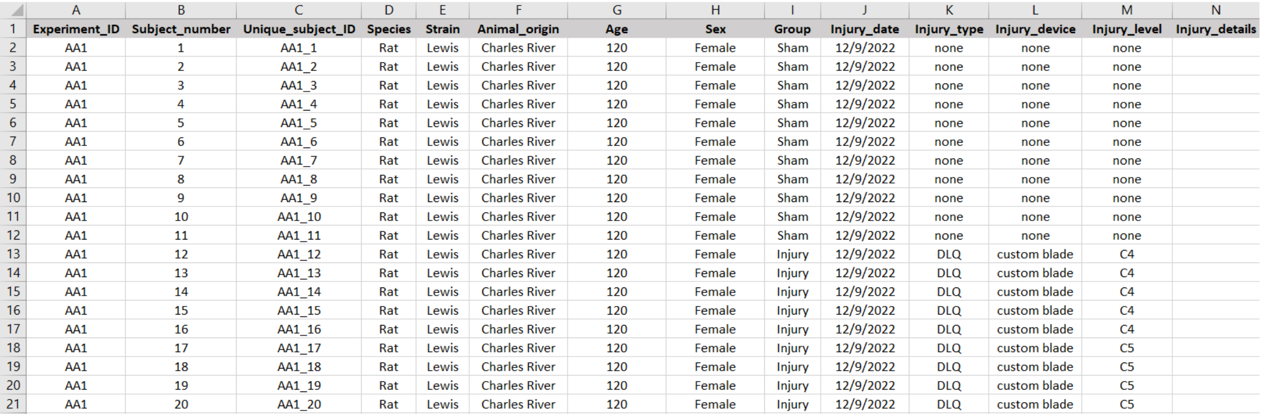
| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Experiment_ID | Date_creation | Investigator | Description | n_subjects | Animal_order_number |
| 2 | AA1 | 12/9/2022 | Abel Torres Espin | Test experiment for demo of catalog | 15 | #85621 (example) |
| 3 | AA2 | 12/10/2022 | Karim Fouad | Testing the hypothesis that drug X do Y | 20 | #85634 (example) |
| 4 | AA3 | 12/11/2022 | Abel Torres Espin | Description for AA3 | 10 | Order number for AA3 |
| 5 | AA4 | 12/12/2022 | Abel Torres Espin | Description for AA4 | 12 | Order number for AA4 |

**Figure 5.** Example of an experiment catalog using a spreadsheet

## 5.3.Subject catalog (Fig. 6).

Each experiment contains a subject list file providing essential information for each subject. This includes experimental variables such as group allocation, species, strain, and any parameters necessary to understand the experiment. Most importantly, it incorporates a unique subject identifier. We suggest tracking at least the information below and adapting the subject list based on your research needs.

- **ExperimentID.** Keeping the experiment identifier in the subject list provides a connection between the experiment and the subject.
- **Subject number.** Most laboratories track subjects by providing a number to the subject.
- **Unique subject identifier.** This is one of the most important pieces of information. Each subject has a unique identifier that does not repeat with other subjects in the lab, past, present, or future. In our example, adding to the experiment identifier (AA1) the subject number (1) is a logical form to track each subject. For example, AA1_1 will identify subject one from experiment AA1.
- **Any other important variable.** The rest of the subject list contains other field-specific variables important for understanding what happens with each subject.



| | A | B | C | D | E | F | G | H | I | J | K | L | M | N |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Experiment_ID | Subject_number | Unique_subject_ID | Species | Strain | Animal_origin | Age | Sex | Group | Injury_date | Injury_type | Injury_device | Injury_level | Injury_details |
| 2 | AA1 | 1 | AA1_1 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 3 | AA1 | 2 | AA1_2 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 4 | AA1 | 3 | AA1_3 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 5 | AA1 | 4 | AA1_4 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 6 | AA1 | 5 | AA1_5 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 7 | AA1 | 6 | AA1_6 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 8 | AA1 | 7 | AA1_7 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 9 | AA1 | 8 | AA1_8 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 10 | AA1 | 9 | AA1_9 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 11 | AA1 | 10 | AA1_10 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 12 | AA1 | 11 | AA1_11 | Rat | Lewis | Charles River | 120 | Female | Sham | 12/9/2022 | none | none | none | |
| 13 | AA1 | 12 | AA1_12 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C4 | |
| 14 | AA1 | 13 | AA1_13 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C4 | |
| 15 | AA1 | 14 | AA1_14 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C4 | |
| 16 | AA1 | 15 | AA1_15 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C4 | |
| 17 | AA1 | 16 | AA1_16 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C4 | |
| 18 | AA1 | 17 | AA1_17 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C5 | |
| 19 | AA1 | 18 | AA1_18 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C5 | |
| 20 | AA1 | 19 | AA1_19 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C5 | |
| 21 | AA1 | 20 | AA1_20 | Rat | Lewis | Charles River | 120 | Female | Injury | 12/9/2022 | DLQ | custom blade | C5 | |

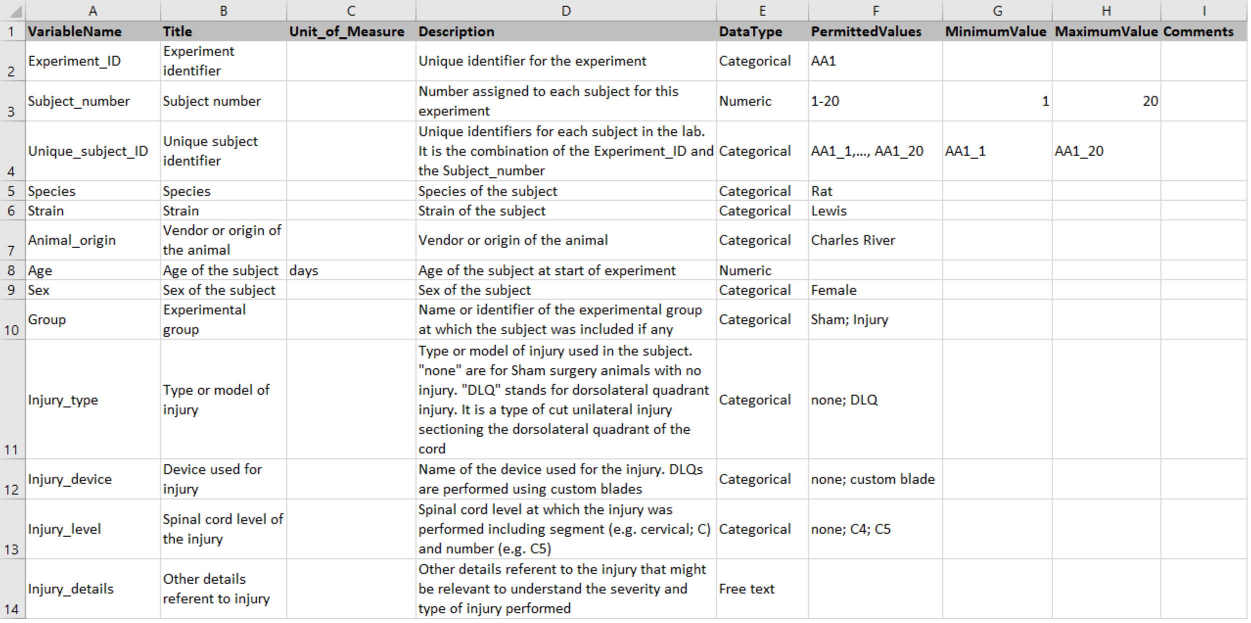
**Figure 6.** Example of a subject catalog using a spreadsheet

## 5.4.Data Dictionary (Fig. 7).

We provided a general description for a data dictionary in the previous section. Here we offer an example, taking the ODC data dictionary specifications, which require the information provided in Table 3.

**Table 3**. Elements of the data dictionary as specified by the ODC standards

| Data dictionary element | Definition |
|---|---|
| **VariableName** | Variables (i.e., column headers) that appear in the dataset. You must include all of your dataset variables in the data dictionary. |
| **Title** | Title is the full name of the variable when the VariableName contains abbreviations or shorthand. If the VariableName is already a complete name, you can copy and paste the VariableName into the Title entry. |
| **Unit_of_Measure** | Units for the variable (if applicable). |
| **Description** | Definitions and descriptions of the variable. The description should explain what the variable represents in enough detail such that a reader can understand the contents of the column in the dataset. |
| **DataType** | Specify whether the variable specifically contains Numeric, Categorical, Ordinal, Date, or Free Text data. |
| **PermittedValues** | If the variable is not numeric or free text, list all possible values here (e.g., "Male, Female" for the variable "Sex"). This field can be left blank if the variable is numeric or free text (use MinimumValue and MaximumValue columns). |
| **MinimumValue** | If the variable is numeric, list the Minimum possible value. For example, if you expect a variable to be between 0-100, write 0 for MinimumValue. If there is no minimum value, leave this blank. |
| **MaximumValue** | If the variable is numeric, list the Maximum possible value. For example, if you expect a variable to be between 0-100, write 100 for MaximumValue. If there is no maximum value, leave this blank. |
| **Comments** | Additional notes such as exclusion criteria, reasons for special values, etc. |



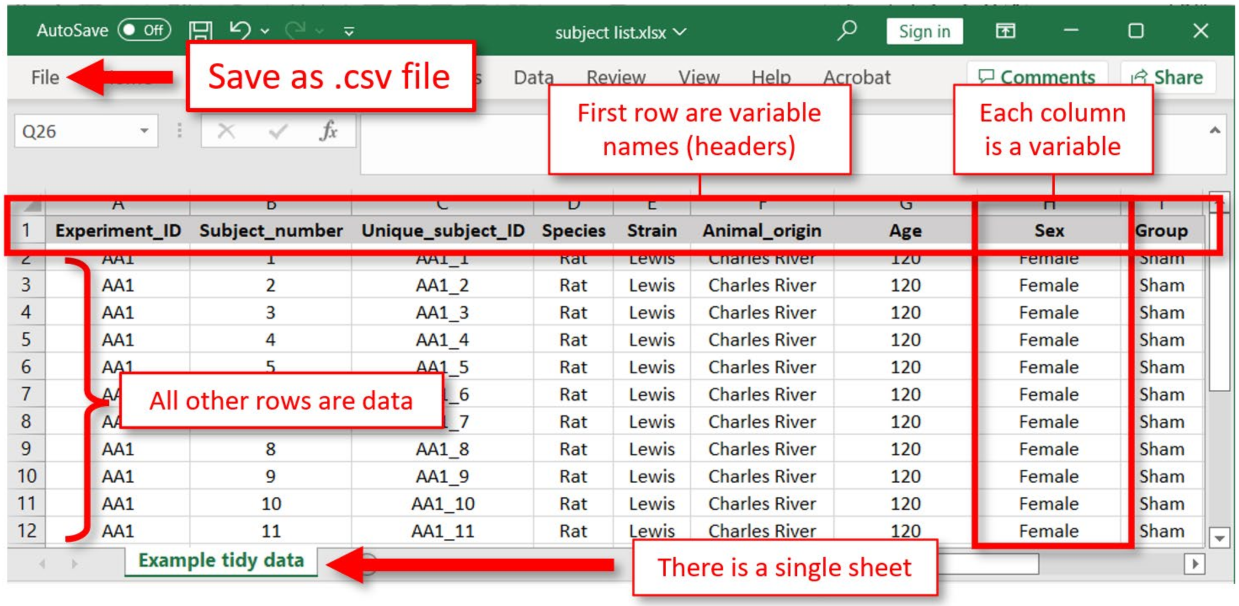
| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | VariableName | Title | Unit_of_Measure | Description | DataType | PermittedValues | MinimumValue | MaximumValue | Comments |
| 2 | Experiment_ID | Experiment identifier | | Unique identifier for the experiment | Categorical | AA1 | | | |
| 3 | Subject_number | Subject number | | Number assigned to each subject for this experiment | Numeric | 1-20 | 1 | 20 | |
| 4 | Unique_subject_ID | Unique subject identifier | | Unique identifiers for each subject in the lab. It is the combination of the Experiment_ID and the Subject_number | Categorical | AA1_1,..., AA1_20 | AA1_1 | AA1_20 | |
| 5 | Species | Species | | Species of the subject | Categorical | Rat | | | |
| 6 | Strain | Strain | | Strain of the subject | Categorical | Lewis | | | |
| 7 | Animal_origin | Vendor or origin of the animal | | Vendor or origin of the animal | Categorical | Charles River | | | |
| 8 | Age | Age of the subject | days | Age of the subject at start of experiment | Numeric | | | | |
| 9 | Sex | Sex of the subject | | Sex of the subject | Categorical | Female | | | |
| 10 | Group | Experimental group | | Name or identifier of the experimental group at which the subject was included if any | Categorical | Sham; Injury | | | |
| 11 | Injury_type | Type or model of injury | | Type or model of injury used in the subject. "none" are for Sham surgery animals with no injury. "DLQ" stands for dorsolateral quadrant injury. It is a type of cut unilateral injury sectioning the dorsolateral quadrant of the cord | Categorical | none; DLQ | | | |
| 12 | Injury_device | Device used for injury | | Name of the device used for the injury. DLQs are performed using custom blades | Categorical | none; custom blade | | | |
| 13 | Injury_level | Spinal cord level of the injury | | Spinal cord level at which the injury was performed including segment (e.g. cervical; C) and number (e.g. C5) | Categorical | none; C4; C5 | | | |
| 14 | Injury_details | Other details referent to injury | | Other details referent to the injury that might be relevant to understand the severity and type of injury performed | Free text | | | | |

**Figure 7.** Example of a data dictionary using a spreadsheet

## 5.5.Processed data files

In biomedical research, we often collect data, which is then processed to extract key variables or metrics. For example, in the field of neurological injuries is common to evaluate different neurological functions or outcomes for each subject over time. The data might be collected through specialized hardware, paper template instruments, video, etc. We suggest organizing all key metrics and variables into spreadsheet files that are easy to manage and share.

- **The tidy data format** (Fig. 8)**.** An excellent way to organize tabular data is using spreadsheets. The tidy data format follows specific rules to keep the data in the spreadsheets clean and easy to read. The tidy format means that the first row is the variable names (aka headers), each subsequent rows are data observations, and each column is a variable. For further information on organizing data in spreadsheets, we recommend (Broman and Woo, 2018; Wickham, 2014).
- **Link to raw data.** If any information in the processed data files comes from specific raw data files such as images or videos, these can be linked by providing the folder and file path.



Save as .csv file

First row are variable names (headers)

Each column is a variable

All other rows are data

There is a single sheet

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Experiment_ID | Subject_number | Unique_subject_ID | Species | Strain | Animal_origin | Age | Sex | Group |
| 2 | AA1 | 1 | AA1_1 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 3 | AA1 | 2 | AA1_2 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 4 | AA1 | 3 | AA1_3 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 5 | AA1 | 4 | AA1_4 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 6 | AA1 | 5 | AA1_5 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 7 | AA | | _6 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 8 | AA | | _7 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 9 | AA1 | 8 | AA1_8 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 10 | AA1 | 9 | AA1_9 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 11 | AA1 | 10 | AA1_10 | Rat | Lewis | Charles River | 120 | Female | Sham |
| 12 | AA1 | 11 | AA1_11 | Rat | Lewis | Charles River | 120 | Female | Sham |

**Figure 8.** Organizing data in a spreadsheet following the tidy format. Although data in this format can be saved in any spreadsheet file format, we recommend .csv files as they are easily interchangeable between systems and readable by a multitude of programs.

# 6. Changing behavior by creating habits

Implementing good data management and sharing workflows within a lab is impossible within a day. There are plenty of challenges for effective data workflow, starting with budget limitations, constant lab member turnover, and the fact that anything to do with data is a field for which most biomedical researchers are not (yet) trained. We have collected some recommendations based on our experience that we hope can help ease these challenges. Our motto is that if you slowly create a habit, the behavior and cultural change will follow.

- **Get tools to help.** Whether it is buying specific software, creating SOPs and templates, or building custom programs, tools can help implement new habits and make them stick. It can also help new lab members to adopt laboratory standards faster.

- **Timely data organization reduces time.** The earlier the data gets organized into standards, the less effort it will take, saving valuable work hours. Waiting to organize data to the end of a project may bring errors from forgotten information. Establishing periods dedicated to managing data regularly can be an efficient way not to have to do it all at once. For example, data collected in notebooks can be entered into the digital data management system at the end of each week.

- **Training.** Most new lab members, especially those with little experience in research, do not have preconceived notions of how data should be collected and managed; they have nothing to unlearn. This means they can be taught good practices and procedures as part of their training. With the increased importance of data in every aspect of research, trainees will significantly benefit from this knowledge. This may include teaching newcomers the SOPs for data management and sharing in the lab, providing learning material on data relevant to the research field, and supporting trainees to attend courses. Most university libraries provide courses on data best practices.

- **Opportunity vs. requirement.** It is all about messaging and encouragement. For example, if a new task is presented as work, it will be seen as such. By pointing out the advantages, it can be seen as a benefit for everybody, a 'win-win' situation. Data mandates can trigger the feeling of more administrative work for compliance. It is important to create an environment where the force of habit encourages good practice. Well-integrated data workflows can seem like a lot of work, but they can increase productivity, as discussed earlier. Once organized, consider a study utilizing historical data to learn about experimental drift and variability or create new hypotheses.

- **Facilitate work supervision.** The data workflows, from data collection to management, analysis, and sharing, can provide natural oversight and productivity monitoring checkpoints. By creating an oversight process for workflow, one can examine data worksheets to monitor productivity, as an expectation for manuscript submission, etc.

# 7. Conclusions

The growing importance of data management and sharing cannot be overstated for researchers, regardless of their training and background. While the task may initially seem overwhelming, understanding the fundamentals and exploring various approaches will reveal the numerous benefits that outweigh the initial investment. Although the specific requirements may vary across laboratories and research areas, our straightforward solution offers customization and scalability to meet individual needs. By embracing data management and sharing practices, researchers can unlock their work's full potential and contribute to advancing their fields.

## References

Almeida, C.A., Torres-Espin, A., Huie, J.R., Sun, D., Noble-Haeusslein, L., Young, W., Beattie, M.S., Bresnahan, J.C., Nielson, J.L., Ferguson, A.R., Submitted. Excavating FAIR Data: The Case of the Multicenter Animal Spinal Cord Injury Study (MASCIS), Blood Pressure, and Neuro-recovery. Neuroinformatics.

Bandrowski, A.E., Martone, M.E., 2016. RRIDs: A Simple Step toward Improving Reproducibility through Rigor and Transparency of Experimental Methods. Neuron 90, 434–436. https://doi.org/10.1016/j.neuron.2016.04.030

Begley, C.G., Ioannidis, J.P.A., 2015. Reproducibility in science: improving the standard for basic and preclinical research. Circ. Res. 116, 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Bierer, B.E., Crosas, M., Pierce, H.H., 2017. Data Authorship as an Incentive to Data Sharing. New England Journal of Medicine 376, 1684–1687. https://doi.org/10.1056/NEJMsb1616595

Biering-Sørensen, F., Alai, S., Anderson, K., Charlifue, S., Chen, Y., DeVivo, M., Flanders, A.E., Jones, L., Kleitman, N., Lans, A., Noonan, V.K., Odenkirchen, J., Steeves, J., Tansey, K., Widerström-Noga, E., Jakeman, L.B., 2015. Common data elements for spinal cord injury clinical research: a National Institute for Neurological Disorders and Stroke project. Spinal Cord 53, 265–277. https://doi.org/10.1038/sc.2014.246

Borgman, C.L., 2015. Big Data, Little Data, No Data: Scholarship in the Networked World. https://doi.org/10.7551/mitpress/9963.001.0001

Boué, S., Byrne, M., Hayes, A.W., Hoeng, J., Peitsch, M.C., 2018. Embracing Transparency Through Data Sharing. Int J Toxicol 37, 466–471. https://doi.org/10.1177/1091581818803880

Broman, K.W., Woo, K.H., 2018. Data Organization in Spreadsheets. The American Statistician 72, 2–10. https://doi.org/10.1080/00031305.2017.1375989

Burke, D.L., Ensor, J., Riley, R.D., 2017. Meta-analysis using individual participant data: one-stage and two-stage approaches, and why they may differ. Stat Med 36, 855–875. https://doi.org/10.1002/sim.7141

Callahan, A., Anderson, K.D., Beattie, M.S., Bixby, J.L., Ferguson, A.R., Fouad, K., Jakeman, L.B., Nielson, J.L., Popovich, P.G., Schwab, J.M., Lemmon, V.P., FAIR Share Workshop Participants, 2017. Developing a data sharing community for spinal cord injury research. Exp. Neurol. 295, 135–143. https://doi.org/10.1016/j.expneurol.2017.05.012

Cameli, M., Novo, G., Tusa, M., Mandoli, G.E., Corrado, G., Benedetto, F., Antonini-Canterin, F., Citro, R., 2018. How to Write a Research Protocol: Tips and Tricks. J Cardiovasc Echogr 28, 151–153. https://doi.org/10.4103/jcecho.jcecho_41_18

Carr, D., Littler, K., 2015. Sharing Research Data to Improve Public Health. J Empir Res Hum Res Ethics 10, 314–316. https://doi.org/10.1177/1556264615593485

Chan, A.-W., Song, F., Vickers, A., Jefferson, T., Dickersin, K., Gøtzsche, P.C., Krumholz, H.M., Ghersi, D., van der Worp, H.B., 2014. Increasing value and reducing waste: addressing inaccessible research. The Lancet 383, 257–266. https://doi.org/10.1016/S0140-6736(13)62296-5

Chervitz, S.A., Deutsch, E.W., Field, D., Parkinson, H., Quackenbush, J., Rocca-Serra, P., Sansone, S.-A., Stoeckert, C.J., Taylor, C.F., Taylor, R., Ball, C.A., 2011. Data standards for Omics data: the basis of data sharing and reuse. Methods Mol Biol 719, 31–69. https://doi.org/10.1007/978-1-61779-027-0_2

Chou, A., Torres-Espín, A., Huie, J.R., Krukowski, K., Lee, S., Nolan, A., Guglielmetti, C., Hawkins, B.E., Chaumeil, M.M., Manley, G.T., Beattie, M.S., Bresnahan, J.C., Martone, M.E., Grethe, J.S., Rosi, S., Ferguson, A.R., 2022. Empowering Data Sharing and Analytics through the Open Data Commons for Traumatic Brain Injury Research. Neurotrauma Rep 3, 139–157. https://doi.org/10.1089/neur.2021.0061

Colavizza, G., Hrynaszkiewicz, I., Staden, I., Whitaker, K., McGillivray, B., 2020. The citation advantage of linking publications to research data. PLoS One 15, e0230416. https://doi.org/10.1371/journal.pone.0230416

Collins, F.S., Tabak, L.A., 2014. NIH plans to enhance reproducibility 2.

Curran, P.J., Hussong, A.M., 2009. Integrative data analysis: the simultaneous analysis of multiple data sets. Psychol Methods 14, 81–100. https://doi.org/10.1037/a0015914

Data sharing is the future, 2023. . Nat Methods 20, 471–471. https://doi.org/10.1038/s41592-023-01865-4

Dempsey, W.P., Foster, I., Fraser, S., Kesselman, C., 2022. Sharing Begins at Home: How Continuous and Ubiquitous FAIRness Can Enhance Research Productivity and Data Reuse. Harvard Data Science Review 4. https://doi.org/10.1162/99608f92.44d21b86

Dhruva, S.S., Ross, J.S., Akar, J.G., Caldwell, B., Childers, K., Chow, W., Ciaccio, L., Coplan, P., Dong, J., Dykhoff, H.J., Johnston, S., Kellogg, T., Long, C., Noseworthy, P.A., Roberts, K., Saha, A., Yoo, A., Shah, N.D., 2020. Aggregating multiple real-world data sources using a patient-centered health-data-sharing platform. NPJ Digit Med 3, 60. https://doi.org/10.1038/s41746-020-0265-z

Fecher, B., Friesike, S., Hebing, M., 2015. What Drives Academic Data Sharing? PLoS One 10, e0118053. https://doi.org/10.1371/journal.pone.0118053

Ferguson, A.R., Irvine, K.-A., Gensel, J.C., Nielson, J.L., Lin, A., Ly, J., Segal, M.R., Ratan, R.R., Bresnahan, J.C., Beattie, M.S., 2013. Derivation of Multivariate Syndromic Outcome Metrics for Consistent Testing across Multiple Models of Cervical Spinal Cord Injury in Rats. PLoS ONE 8, e59712. https://doi.org/10.1371/journal.pone.0059712

Ferguson, A.R., Nielson, J.L., Cragin, M.H., Bandrowski, A.E., Martone, M.E., 2014. Big data from small data: data-sharing in the “long tail” of neuroscience. Nature Neuroscience 17, 1442–1447. https://doi.org/10.1038/nn.3838

Fire, M., Guestrin, C., 2019. Over-optimization of academic publishing metrics: observing Goodhart’s Law in action. GigaScience 8, giz053. https://doi.org/10.1093/gigascience/giz053

Flanagin, A., Curfman, G., Bibbins-Domingo, K., 2022. Data Sharing and the Growth of Medical Knowledge. JAMA 328, 2398–2399. https://doi.org/10.1001/jama.2022.22837

Fouad, K., Bixby, J.L., Callahan, A., Grethe, J.S., Jakeman, L.B., Lemmon, V.P., Magnuson, D.S., Martone, M.E., Nielson, J.L., Schwab, J., Taylor-Burds, C., Tetzlaff, W., Torres-Espín, A., Ferguson, A.R., 2019. FAIR SCI Ahead: the evolution of the Open Data Commons for preclinical spinal cord injury research (ODC-SCI.org). J. Neurotrauma. https://doi.org/10.1089/neu.2019.6674

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y., Zhang, J., 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biology 5, R80. https://doi.org/10.1186/gb-2004-5-10-r80

Gorgolewski, K.J., Margulies, D.S., Milham, M.P., 2013. Making Data Sharing Count: A Publication-Based Solution. Front. Neurosci. 7. https://doi.org/10.3389/fnins.2013.00009

Government of Canada, I., 2021. Tri-Agency Statement of Principles on Digital Data Management [WWW Document]. URL https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-statement-principles-digital-data-management (accessed 5.31.23).

Halchenko, Y.O., Meyer, K., Poldrack, B., Solanky, D.S., Wagner, A.S., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S.S., Mönch, C., Markiewicz, C.J., Waite, L., Shlyakhter, I., Vega, A. de la, Hayashi, S., Häusler, C.O., Poline, J.-B., Kadelka, T., Skytén, K., Jarecka, D., Kennedy, D., Strauss, T., Cieslak, M., Vavra, P., Ioanas, H.-I., Schneider, R., Pflüger, M., Haxby, J.V., Eickhoff, S.B., Hanke, M., 2021. DataLad: distributed system for joint management of code, data, and their relationship. Journal of Open Source Software 6, 3262. https://doi.org/10.21105/joss.03262

Hicks, R., Giacino, J., Harrison-Felix, C., Manley, G., Valadka, A., Wilde, E.A., 2013. Progress in Developing Common Data Elements for Traumatic Brain Injury Research: Version Two – The End of the Beginning. Journal of Neurotrauma 30, 1852–1861. https://doi.org/10.1089/neu.2013.2938

Hu, Q., Hutson, A., Liu, S., Morgan, M., Liu, Q., 2021. Bioconductor toolchain for reproducible bioinformatics pipelines using Rcwl and RcwlPipelines. Bioinformatics 37, 3351–3352. https://doi.org/10.1093/bioinformatics/btab208

Hu, S., Wu, G., Wu, B., Du, Z., Zhang, Y., 2022. Rehabilitative training paired with peripheral stimulation promotes motor recovery after ischemic cerebral stroke. Exp Neurol 349, 113960. https://doi.org/10.1016/j.expneurol.2021.113960

Huser, V., Sastry, C., Breymaier, M., Idriss, A., Cimino, J.J., 2015. Standardizing data exchange for clinical research protocols and case report forms: An assessment of the suitability of the Clinical Data Interchange Standards Consortium (CDISC) Operational Data Model (ODM). J Biomed Inform 57, 88–99. https://doi.org/10.1016/j.jbi.2015.06.023

Ioannidis, J.P.A., 2014. How to Make More Published Research True. PLoS Med 11, e1001747. https://doi.org/10.1371/journal.pmed.1001747

Karpen, S.R., White, J.K., Mullin, A.P., O'Doherty, I., Hudson, L.D., Romero, K., Sivakumaran, S., Stephenson, D., Turner, E.C., Larkindale, J., 2021. Effective Data Sharing as a Conduit for Advancing Medical Product Development. Ther Innov Regul Sci 55, 591–600. https://doi.org/10.1007/s43441-020-00255-8

Kennedy, D.N., 2012. The Benefits of Preparing Data for Sharing Even When You Don't. Neuroinform 10, 223–224. https://doi.org/10.1007/s12021-012-9154-1

LaPlaca, M.C., Huie, J.R., Alam, H.B., Bachstetter, A.D., Bayir, H., Bellgowan, P.F., Cummings, D., Dixon, C.E., Ferguson, A.R., Ferland-Beckham, C., Floyd, C.L., Friess, S.H., Galanopoulou, A.S., Hall, E.D., Harris, N.G., Hawkins, B.E., Hicks, R.R., Hulbert, L.E., Johnson, V.E., Kabitzke, P.A., Lafrenaye, A.D., Lemmon, V.P., Lifshitz, C.W., Lifshitz, J., Loane, D.J., Misquitta, L., Nikolian, V.C., Noble-Haeusslein, L.J., Smith, D.H., Taylor-Burds, C., Umoh, N., Vovk, O., Williams, A.M., Young, M., Zai, L.J., 2021. Pre-Clinical Common Data Elements for Traumatic Brain Injury Research: Progress and Use Cases. J Neurotrauma 38, 1399–1410. https://doi.org/10.1089/neu.2020.7328

Lee, J.E., Sung, J.H., Barnett, M.E., Norris, K., 2016. User-Friendly Data-Sharing Practices for Fostering Collaboration within a Research Network: Roles of a Vanguard Center for a Community-Based Study. International Journal of Environmental Research and Public Health 13, 34. https://doi.org/10.3390/ijerph13010034

Leonelli, S., 2015. What Counts as Scientific Data? A Relational Framework. Philos Sci 82, 810–821.

Levesque, R.J.R., 2017. Data Sharing Mandates, Developmental Science, and Responsibly Supporting Authors. J Youth Adolescence 46, 2401–2406. https://doi.org/10.1007/s10964-017-0741-1

Manley, G.T., Diaz-Arrastia, R., Brophy, M., Engel, D., Goodman, C., Gwinn, K., Veenstra, T.D., Ling, G., Ottens, A.K., Tortella, F., Hayes, R.L., 2010. Common data elements for traumatic brain injury: recommendations from the biospecimens and biomarkers working group. Arch Phys Med Rehabil 91, 1667–1672. https://doi.org/10.1016/j.apmr.2010.05.018

Markiewicz, C.J., Gorgolewski, K.J., Feingold, F., Blair, R., Halchenko, Y.O., Miller, E., Hardcastle, N., Wexler, J., Esteban, O., Goncavles, M., Jwa, A., Poldrack, R., 2023. The OpenNeuro resource for sharing of neuroscience data. eLife 10, e71774. https://doi.org/10.7554/eLife.71774

Martone, M.E., Garcia-Castro, A., VandenBos, G.R., 2018. Data Sharing in Psychology. Am Psychol 73, 111–125. https://doi.org/10.1037/amp0000242

Martone, M.E., Nakamura, R., 2022. Changing the Culture on Data Management and Sharing: Getting Ready for the New NIH Data Sharing Policy. Harvard Data Science Review. https://doi.org/10.1162/99608f92.6650ce2b

Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., Ghosh, S., Glatard, T., Gorgolewski, K.J., Helmer, K.G., Jenkinson, M., Keator, D.B., Nichols, B.N., Poline, J.-B., Reynolds, R., Sochat, V., Turner, J., Nichols, T.E., 2016. Sharing brain mapping statistical results with the neuroimaging data model. Sci Data 3, 160102. https://doi.org/10.1038/sdata.2016.102

Nielson, J.L., Paquette, J., Liu, A.W., Guandique, C.F., Tovar, C.A., Inoue, T., Irvine, K.-A., Gensel, J.C., Kloke, J., Petrossian, T.C., Lum, P.Y., Carlsson, G.E., Manley, G.T., Young, W., Beattie, M.S., Bresnahan, J.C., Ferguson, A.R., 2015. Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. Nat Commun 6, 8581. https://doi.org/10.1038/ncomms9581

NIH DMS policy, 2023. Data Management & Sharing Policy Overview | Data Sharing [WWW Document]. URL https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policies/data-management-and-sharing-policy-overview (accessed 5.31.23).

Niso, G., Gorgolewski, K.J., Bock, E., Brooks, T.L., Flandin, G., Gramfort, A., Henson, R.N., Jas, M., Litvak, V., T. Moreau, J., Oostenveld, R., Schoffelen, J.-M., Tadel, F., Wexler, J., Baillet, S., 2018. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. Sci Data 5, 180110. https://doi.org/10.1038/sdata.2018.110

OECD, 2006. Recommendation of the Council concerning Access to Research Data from Public Funding [WWW Document]. URL https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0347 (accessed 5.31.23).

Ohmann, C., Banzi, R., Canham, S., Battaglia, S., Matei, M., Ariyo, C., Becnel, L., Bierer, B., Bowers, S., Clivio, L., Dias, M., Druml, C., Faure, H., Fenner, M., Galvez, J., Ghersi, D., Gluud, C., Groves, T., Houston, P., Karam, G., Kalra, D., Knowles, R.L., Krleža-Jerić, K., Kubiak, C., Kuchinke, W., Kush, R., Lukkarinen, A., Marques, P.S., Newbigging, A., O'Callaghan, J., Ravaud, P., Schlünder, I., Shanahan, D., Sitter, H., Spalding, D., Tudur-Smith, C., Reusel, P. van, Veen, E.-B. van, Visser, G.R., Wilson, J., Demotes-Mainard, J., 2017. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open 7, e018647. https://doi.org/10.1136/bmjopen-2017-018647

Riley, R.D., Lambert, P.C., Abo-Zaid, G., 2010. Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 340. https://doi.org/10.1136/bmj.c221

Roundtable on Environmental Health Sciences, R., Practice, B. on P.H. and P.H., Division, H. and M., National Academies of Sciences, E., 2016. The Benefits of Data Sharing, Principles and Obstacles for Sharing Data from Environmental Health Research: Workshop Summary. National Academies Press (US).

Rübel, O., Dougherty, M., Prabhat,  null, Denes, P., Conant, D., Chang, E.F., Bouchard, K., 2016. Methods for Specifying Scientific Data Standards and Modeling Relationships with Applications to Neuroscience. Front Neuroinform 10, 48. https://doi.org/10.3389/fninf.2016.00048

Scargle, J.D., 1999. Publication Bias (The "File-Drawer Problem") in Scientific Inference. arXiv:physics/9909033.

Schembera, B., Durán, J.M., 2020. Dark Data as the New Challenge for Big Data Science and the Introduction of the Scientific Data Officer. Philos. Technol. 33, 93–115. https://doi.org/10.1007/s13347-019-00346-x

Sherpa Juliet, 2023. Research Funders' Open Access Policies - Sherpa Services [WWW Document]. URL https://v2.sherpa.ac.uk/juliet/ (accessed 5.31.23).

Smith, D.H., Hicks, R.R., Johnson, V.E., Bergstrom, D.A., Cummings, D.M., Noble, L.J., Hovda, D., Whalen, M., Ahlers, S.T., LaPlaca, M., Tortella, F.C., Duhaime, A.-C., Dixon, C.E., 2015. Pre-Clinical Traumatic Brain Injury Common Data Elements: Toward a Common Language Across Laboratories. Journal of Neurotrauma 32, 1725–1735. https://doi.org/10.1089/neu.2014.3861

Sperr, E., 2016. PubMed by Year [WWW Document]. URL http://esperr.github.io/pubmed-by-year/ (accessed 5.31.23).

Teytelman, L., Stoliartchouk, A., Kindler, L., Hurwitz, B.L., 2016. Protocols.io: Virtual Communities for Protocol Development and Discussion. PLOS Biology 14, e1002538. https://doi.org/10.1371/journal.pbio.1002538

Thompson, H.J., Vavilala, M.S., Rivara, F.P., 2015. Common Data Elements and Federal Interagency Traumatic Brain Injury Research Informatics System for TBI Research. Annu Rev Nurs Res 33, 1–11. https://doi.org/10.1891/0739-6686.33.1

Torres-Espín, A., 2023. ATE-DRIVEN-lab/practical_guide_DMS: Practical guide DMS v1.0. https://doi.org/10.5281/zenodo.8071997

Torres-Espín, A., Almeida, C.A., Chou, A., Huie, J.R., Chiu, M., Vavrek, R., Sacramento, J., Orr, M.B., Gensel, J.C., Grethe, J.S., Martone, M.E., Fouad, K., Ferguson, A.R., STREET-FAIR Workshop Participants, 2021. Promoting FAIR Data Through Community-driven Agile Design: the Open Data Commons for Spinal Cord Injury (odc-sci.org). Neuroinformatics. https://doi.org/10.1007/s12021-021-09533-8

van der Steen, J.T., Kruse, R.L., Szafara, K.L., Mehr, D.R., van der Wal, G., Ribbe, M.W., D’Agostino, R.B., 2008. Benefits and pitfalls of pooling datasets from comparable observational studies: combining US and Dutch nursing home studies. Palliat Med 22, 750–759. https://doi.org/10.1177/0269216308094102

Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A., 2011. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39, W541-545. https://doi.org/10.1093/nar/gkr469

Whyte, J., Vasterling, J., Manley, G.T., 2010. Common data elements for research on traumatic brain injury and psychological health: current status and future development. Arch Phys Med Rehabil 91, 1692–1696. https://doi.org/10.1016/j.apmr.2010.06.031

Wickham, H., 2014. Tidy Data. Journal of Statistical Software 59, 1–23. https://doi.org/10.18637/jss.v059.i10

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., ’t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18