

Lawrence Berkeley National Laboratory

LBL Publications

Title

Ten simple rules for managing laboratory information

Permalink

<https://escholarship.org/uc/item/4nx3j9jt>

Journal

PLOS Computational Biology, 19(12)

ISSN

1553-734X

Authors

Berezin, Casey-Tyler

Aguilera, Luis U

Billerbeck, Sonja

et al.

Publication Date

2023

DOI

10.1371/journal.pcbi.1011652

Peer reviewed

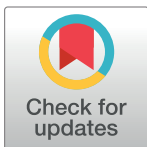
EDUCATION

Ten simple rules for managing laboratory information

Casey-Tyler Berezin¹, Luis U. Aguilera¹, Sonja Billerbeck², Philip E. Bourne^{3,4}, Douglas Densmore⁵, Paul Freemont⁶, Thomas E. Goroehowski^{7,8}, Sarah I. Hernandez¹, Nathan J. Hillson^{9,10,11}, Connor R. King¹, Michael Köpke¹², Shuyi Ma¹³, Katie M. Miller¹, Tae Seok Moon¹⁴, Jason H. Moore¹⁵, Brian Munsky¹, Chris J. Myers¹⁶, Dequina A. Nicholas¹⁷, Samuel J. Peccoud¹⁸, Wen Zhou¹⁹, Jean Peccoud^{1*}

1 Department of Chemical and Biological Engineering, Colorado State University, Fort Collins, Colorado, United States of America, **2** Molecular Microbiology Unit, Faculty of Science and Engineering, University of Groningen, Groningen, the Netherlands, **3** School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America, **4** Department of Biomedical Engineering, University of Virginia, Charlottesville, Virginia, United States of America, **5** College of Engineering, Boston University, Boston, Massachusetts, United States of America, **6** Department of Infectious Disease, Imperial College, London, United Kingdom, **7** School of Biological Sciences, University of Bristol, Bristol, United Kingdom, **8** BrisEngBio, University of Bristol, Bristol, United Kingdom, **9** Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America, **10** US Department of Energy Agile BioFoundry, Emeryville, California, United States of America, **11** US Department of Energy Joint BioEnergy Institute, Emeryville, California, United States of America, **12** LanzaTech, Skokie, Illinois, United States of America, **13** Center for Global Infectious Disease Research, Seattle Children's Hospital, University of Washington Medicine, Seattle, Washington, United States of America, **14** Department of Energy, Environmental & Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri, United States of America, **15** Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, California, United States of America, **16** Department of Electrical, Computer & Energy Engineering, University of Colorado Boulder, Boulder, Colorado, United States of America, **17** Department of Molecular Biology & Biochemistry, University of California Irvine, Irvine, California, United States of America, **18** Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, Colorado, United States of America, **19** Department of Statistics, Colorado State University, Fort Collins, Colorado, United States of America

* jean.peccoud@colostate.edu



OPEN ACCESS

Citation: Berezin C-T, Aguilera LU, Billerbeck S, Bourne PE, Densmore D, Freemont P, et al. (2023) Ten simple rules for managing laboratory information. *PLoS Comput Biol* 19(12): e1011652. <https://doi.org/10.1371/journal.pcbi.1011652>

Editor: Scott Markel, Dassault Systemes BIOVIA, UNITED STATES

Published: December 7, 2023

Copyright: This is an open access article, free of all copyright, and may be freely reproduced, distributed, transmitted, modified, built upon, or otherwise used by anyone for any lawful purpose. The work is made available under the [Creative Commons CC0](https://creativecommons.org/licenses/by/4.0/) public domain dedication.

Funding: J.P., S.H., C.K., K.M., and C.-T.B. are supported by the National Science Foundation (award #2123367) and the National Institutes of Health (R01GM147816, T32GM132057). T.S.M. is supported by the Defense Advanced Research Projects Agency (N660012324032), the Office of Naval Research (N00014-21-1-2206), U.S. Environmental Protection Agency (84020501), the National Institutes of Health (R01 AT009741), and the National Science Foundation (MCB-2001743 and EF-2222403). C.J.M. is supported by the National Science Foundation (MCB-2231864) and the National Institutes Standards and Technology (70NANB21H103). D.A.N. is supported by the National Institutes of Health (R00HD098330 and DP2AI171121). B.M. and L.U.A. are supported by

Author summary

Information is the cornerstone of research, from experimental (meta)data and computational processes to complex inventories of reagents and equipment. These 10 simple rules discuss best practices for leveraging laboratory information management systems to transform this large information load into useful scientific findings.

Introduction

The development of mathematical models that can predict the properties of biological systems is the holy grail of computational biology [1,2]. Such models can be used to test biological hypotheses [3], quantify the risk of developing diseases [3], guide the development of biomanufactured products [4], engineer new systems meeting user-defined specifications, and much more [4,5]. Irrespective of a model's application and the conceptual framework used to build it, the modeling process proceeds through a common iterative workflow. A model is first evaluated by fitting its parameters such that its behavior matches experimental data. Models that fit previous observations are then further validated by comparing the model predictions with the results of new observations that are outside the scope of the initial data set (Fig 1).

the National Institutes of Health (R35 GM124747). T.E.G. was supported by a Royal Society University Research Fellowship grant UF160357, a Turing Fellowship from The Alan Turing Institute under the EPSRC grant EP/N510129/1. W.Z. is supported by the National Institutes of Health (R01GM144961) and the National Science Foundation (IOS1922701). This work was part of the Agile BioFoundry (agilebiofoundry.org) supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, and was part of the DOE Joint BioEnergy Institute (jbei.org) supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility or the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed or represents that its use would not infringe privately owned rights. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (<http://energy.gov/downloads/doe-public-access-plan>). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: J.P., S.P., and K.M. have a financial interest in GenoFAB, Inc., M.K. is an employee of LanzaTech. N.J.H. has a financial interest in TeselaGen Biotechnology, Inc. and Ansa Biotechnologies, Inc. GenoFAB Inc. and TeselaGen Biotechnology, Inc. provide research information management systems. These companies may benefit or be perceived as benefiting from this publication.

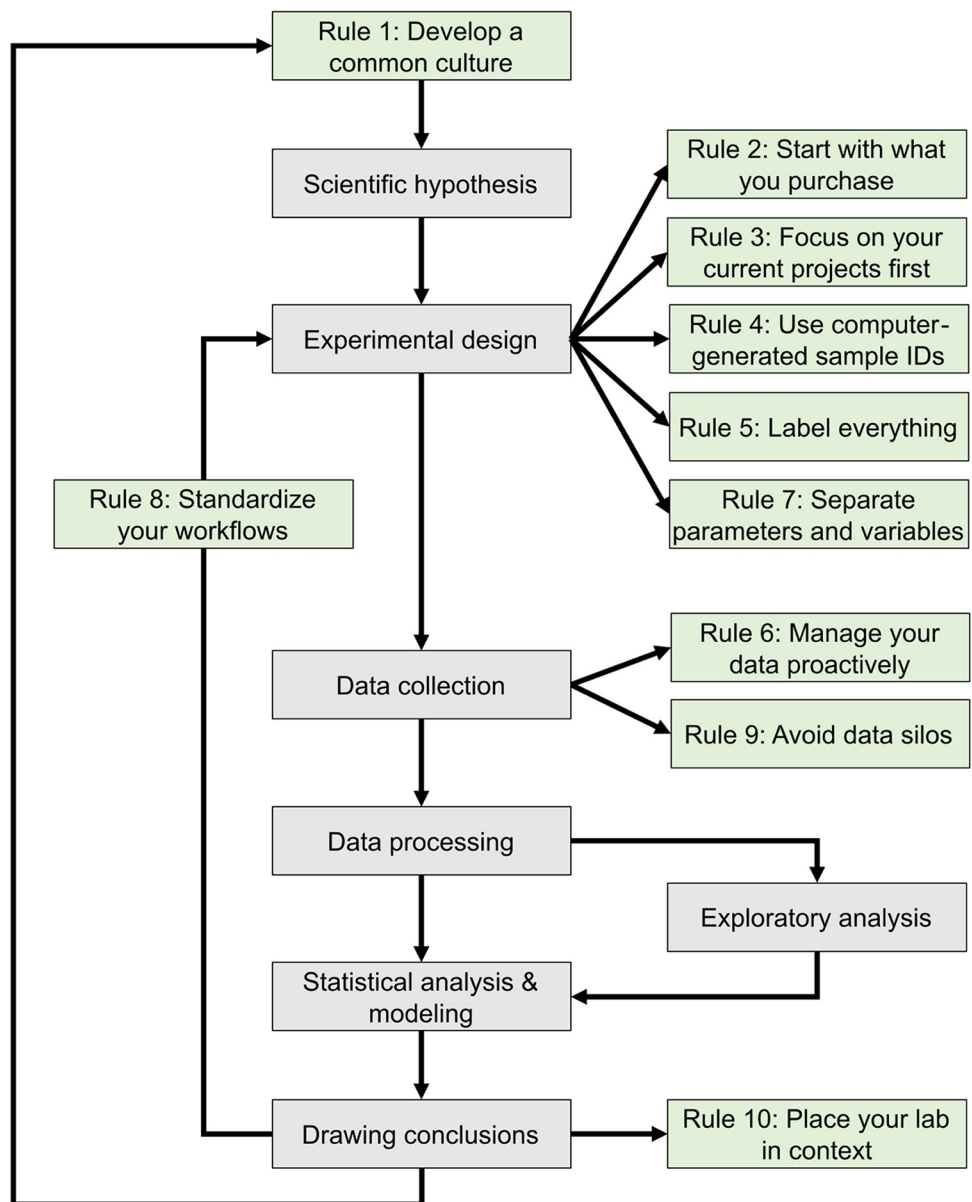


Fig 1. Information management enhances the experimental and modeling cycle. Our 10 simple rules for managing laboratory information (green) augment the cycle of hypothesis formulation, design, data analysis, modeling, and decision-making (gray). The experimental design phase is improved by carefully tracking your inventory, samples, parameters, and variables. Proactive data management and the thoughtful use of databases facilitate statistical and exploratory analyses as well as the development of conclusions that inform the next round of experiments. Frequent reevaluation of project, team, and workflow success is a critical component of refining experimental processes, developing a common culture, and positioning your research group in the greater scientific context.

<https://doi.org/10.1371/journal.pcbi.1011652.g001>

Historically, the collection of experimental data and the development of mathematical models were performed by different scientific communities [6]. Computational biologists had little control over the nature and quality of the data they could access. With the emergence of systems biology and synthetic biology, the boundary between experimental and computational biology has become increasingly blurred [6]. Many labs and junior scientists now have expertise in both producing and analyzing large volumes of digital data produced by high-

throughput workflows and an ever-expanding collection of digital instruments [7]. In this context, it is critically important to properly organize the exponentially growing volumes of experimental data to ensure they can support the development of models that can guide the next round of experiments [8].

We are a group of scientists representing a broad range of scientific specialties, from clinical research to industrial biotechnology. Collectively, we have expertise in experimental biology, data science, and mathematical modeling. Some of us work in academia, while others work in industry. We have all faced the challenges of keeping track of laboratory operations to produce high-quality data suitable for analysis. We have experience using a variety of tools, including spreadsheets, open-source software, homegrown databases, and commercial solutions to manage our data. Irreproducible experiments, projects that failed to meet their goals, datasets we collected but never managed to analyze, and freezers full of unusable samples have taught us the hard way lessons that have led to these 10 simple rules for managing laboratory information.

This journal has published several sets of rules regarding best practices in overall research design [9,10], as well as the computational parts of research workflows, including data management [11–13] and software development practices [14–16]. The purpose of these 10 rules (Fig 1) is to guide the development and configuration of lab information management systems (LIMS). LIMS typically offer lab notebook, inventory, workflow planning, and data management features, allowing users to connect data production and data analysis to ensure that useful information can be extracted from experimental data and increase reproducibility [17,18]. These rules can also be used to develop training programs and lab management policies. Although we all agree that applying these rules increases the value of the data we produce in our laboratories, we also acknowledge that enforcing them is challenging. It relies on the successful integration of effective software tools, training programs, lab management policies, and the will to abide by these policies. Each lab must find the most effective way to adopt these rules to suit their unique environment.

Rule 1: Develop a common culture

Data-driven research projects generally require contributions from multiple stakeholders with complementary expertise. The project's success depends on the entire team developing a common vision of the project objectives and the approaches to be used [19–21]. Interdisciplinary teams, in particular, must establish a common language as well as mutual expectations for experimental and publication timelines [19]. Unless the team develops a common culture, one stakeholder group can drive the project and impose its vision on the other groups. Although interdisciplinary (i.e., wet-lab and computational) training is becoming more common in academia, it is not unusual for experimentalists to regard data analysis as a technique they can acquire simply by hiring a student with computer programming skills. In a corporate environment, research informatics is often part of the information technology group whose mission is to support scientists who drive the research agenda. In both situations, the research agenda is driven by stakeholders who are unlikely to produce the most usable datasets because they lack sufficient understanding of data modeling [20]. Perhaps less frequently, there is also the situation where the research agenda is driven by people with expertise in data analysis. Because they may not appreciate the subtleties of experimental methods, they may find it difficult to engage experimentalists in collaborations aimed at testing their models [20]. Alternatively, their research may be limited to the analysis of disparate sets of previously published datasets [19]. Thus, interdisciplinary collaboration is key to maximizing the insights you gain from your data.

The development of a common culture, within a single laboratory or across interdisciplinary research teams, must begin with a thorough onboarding process for each member regarding general lab procedures, research goals, and individual responsibilities and expectations [21,22]. Implementing a LIMS requires perseverance by users, thus a major determinant of the success of a LIMS is whether end-users are involved in the development process [17,23]. When the input and suggestions of end-users are considered, they are more likely to engage with and upkeep the LIMS on a daily basis [23]. The long-term success of research endeavors then requires continued training and reevaluation of project goals and success [19,21] (Fig 1).

These 10 simple rules apply to transdisciplinary teams that have developed a common culture allowing experimentalists to gain a basic understanding of the modeling process and modelers to have some familiarity with the experimental processes generating the data they will analyze [19]. Teams that lack a common vision of data-driven research are encouraged to work toward acquiring this common vision through frequent communication and mutual goal setting [19,20]. Discussing these 10 simple rules in group meetings may aid in initiating this process.

Rule 2: Start with what you purchase

All the data produced by your lab are derived from things you have purchased, including supplies (consumables), equipment, and contract manufactured reagents, such as oligonucleotides or synthetic genes. In many cases, (meta)data on items in your inventory may be just as important as experimentally derived data, and as such, should be managed according to the Findability, Accessibility, Interoperability, and Reuse (FAIR) principles for (meta)data management (<https://www.go-fair.org/fair-principles/>) [24]. Assembling an inventory of supplies and equipment with their associated locations can be handled in a few weeks by junior personnel without major interruption of laboratory operations, although establishing a thorough inventory may be more difficult and time-consuming for smaller labs with fewer members. Nevertheless, managing your lab inventory provides an immediate return on investment by positively impacting laboratory operations in several ways. People can quickly find the supplies and equipment they need to work, supplies are ordered with appropriate advance notice to minimize work stoppage, and data variation is reduced due to standardized supplies and the ability to track lot numbers easily [17,25,26] (Fig 1).

Many labs still use Excel to keep track of inventory despite the existence of several more sophisticated databases and LIMS (e.g., [Benchling](#), [Quartzy](#), [GenoFAB](#), [LabWare](#), [LabVantage](#), [TeselaGen](#)) [25]. These can facilitate real-time inventory tracking unlike a static document, increasing the Findability and Accessibility of inventory data. While some systems are specialized for certain types of inventories (e.g., animal colonies or frozen reagents), others are capable of tracking any type of reagent or item imaginable [25]. When considering what items to keep track of, there are 3 main considerations: expiration, maintenance, and ease of access.

Most labs manage their supplies through periodic cleanups of the lab, during which they sort through freezers, chemical cabinets, and other storage areas; review their contents; and dispose of supplies that are past their expiration date or are no longer useful. By actively tracking expiration dates and reagent use in a LIMS, you can decrease the frequency of such cleanups since the LIMS will alert users when expiration dates are approaching or when supplies are running low. This can prevent costly items from being wasted because they are expired or forgotten, and furthermore, the cost of products can be tracked and used to inform which experiments are performed.

LIMS can also support the use and service of key laboratory equipment. User manuals, service dates, warranties, and other identifying information can be attached directly to the

equipment record, which allows for timely service and maintenance of the equipment. Adding equipment to the inventory can also prevent accidental losses in shared spaces where it is easy for people to borrow equipment without returning it. The label attached to the equipment (Rule 5) acts as an indication of ownership that limits the risk of ownership confusion when almost identical pieces of equipment are owned by neighboring laboratories. As the laboratory inventory should focus on larger, more expensive equipment and supplies, inexpensive and easily obtained equipment (i.e., office supplies) may not need to be inventoried. An additional benefit of inventory management in a LIMS is the ability to create a record connecting specific equipment and supplies to specific people and projects, which can be used to detect potential sources of technical bias and variability (Rules 4 and 5).

Rule 3: Focus on your current projects first

After establishing an inventory of supplies and equipment, it is natural to consider using a similar approach with the samples that have accumulated over the years in freezers or other storage locations. This can be overwhelming because the number of samples will be orders of magnitude larger than the number of supplies. In addition, documenting them is likely to require more effort than simply retrieving a product documentation from a vendor's catalog.

Allocating limited resources to making an inventory of samples generated by past projects may not benefit current and future projects. A more practical approach is to prioritize tracking samples generated by ongoing projects and document samples generated by past projects on an as-needed basis.

Inventory your samples before you generate them

It is a common mistake to create sample records well after they were produced in the lab. The risks of this retroactive approach to recordkeeping include information loss, as well as selective recordkeeping, in which only some samples are deemed important enough to document while most temporary samples are not, even though they may provide critical information.

A more proactive approach avoids these pitfalls. When somebody walks into a lab to start an experiment, the samples that will be generated by this experiment should be known. It is possible to create the computer records corresponding to these samples before initiating the laboratory processes that generates the physical samples. The creation of a sample record can therefore be seen as part of the experiment planning process (Fig 1). This makes it possible to preemptively print labels that will be used to track samples used at different stages of the process (Rule 5).

It may also be useful to assign statuses to samples as they progress through different stages of their life cycle, such as "to do," "in progress," "completed," or "canceled," to differentiate active works in progress from the backlog and samples generated by previous experiments. As the experimental process moves forward, data can be continually appended to the sample computer record. For example, the field to capture the concentration of a solution would be filled after the solution has been prepared. Thus, the success, or failure, of experiments can be easily documented and used to inform the next round of experiments.

Develop sample retention policies

It is always unpleasant to have to throw away poorly documented samples. The best strategy to avoid this outcome is to develop policies to discard only samples that will not be used in the future, a process rendered more objective and straightforward with adequate documentation. Properly structured workflows (Rule 8) should define criteria for which samples should be

kept and for how long. All lab members should be trained in these policies to ensure consistency, and policies should be revisited as new research operating procedures are initiated.

It can be tempting to keep every tube or plate that still contains something as a backup. This conservative strategy generates clutter, confusion, and reproducibility issues, especially in the absence of documentation. While it makes sense to keep some intermediates during the execution of a complex experimental workflow, the successful completion of the experiment should trigger the elimination of intermediates that have lost their purpose, have limited shelf life, and/or are not reusable. During this intermediate step, samples that are deemed as critical backup should be stored in a separate location from the working sample to minimize the likelihood of loss of both samples in case of electrical failure, fire, etc. Using clear labels (Rules 4 and 5) and storing intermediate samples in dedicated storage locations can help with the enforcement of sample disposal policies.

Rule 4: Use computer-generated sample identification numbers

Generating sample names is probably not the best use of scientists' creativity. Many labs still rely on manually generated sample names that may look something like "JP PCR 8/23 4." Manually generated sample names are time-consuming to generate, difficult to interpret, and often contain insufficient information. Therefore, they should not be the primary identifier used to track samples.

Instead, computer-generated sample identification numbers (Sample ID) should be utilized as the primary ID as they are able to overcome these limitations. Rather than describing the sample, a computer-generated sample ID provides a link between a physical sample and a database entry that contains more information associated with the sample. The Sample ID is the only piece of information that needs to be printed on the sample label (Rule 5) because it allows researchers to retrieve all the sample information from a database. A sample tracking system should rely on both computer-readable and human-readable Sample IDs.

Computer-readable IDs

Since the primary purpose of a sample ID is to provide a link between a physical sample and the computer record that describes the sample, it saves time to rely on Sample IDs that can be scanned by a reader or even a smartphone [27,28] (Fig 2). Barcodes are special fonts to print data in a format that can be easily read by an optical sensor [29]. There are also newer alternatives, such as quick response (QR) codes, data matrices, or radio-frequency identification (RFID), to tag samples [30,31]. QR codes and data matrices are 2D barcodes that are cheaper to generate than RFID tags and store more data than traditional barcodes [27]. Nevertheless, these technologies encode a key that points to a database record.

Uniqueness is the most important property of the data encoded in barcodes, and the use of unique and persistent identifiers is a critical component of the Findability of your (meta)data [24]. Several vendors now offer products with 2D barcodes printed on the side or bottom of the tube. It is common for such products, as well as lab-made reagents, to be split across multiple tubes or moved from one tube to another. In these cases, each of these "new" samples should have unique barcodes. A barcoding system can therefore facilitate the accurate identification of "parent" samples (e.g., a stock solution with ID X) and the unique "child" samples derived from them (e.g., aliquots of the stock solution with IDs Y and Z).

Human-readable IDs

While computer-readable IDs should be the main ID used when tracking a sample or supply, it is sometimes necessary for laboratory personnel to have a secondary sample ID they can

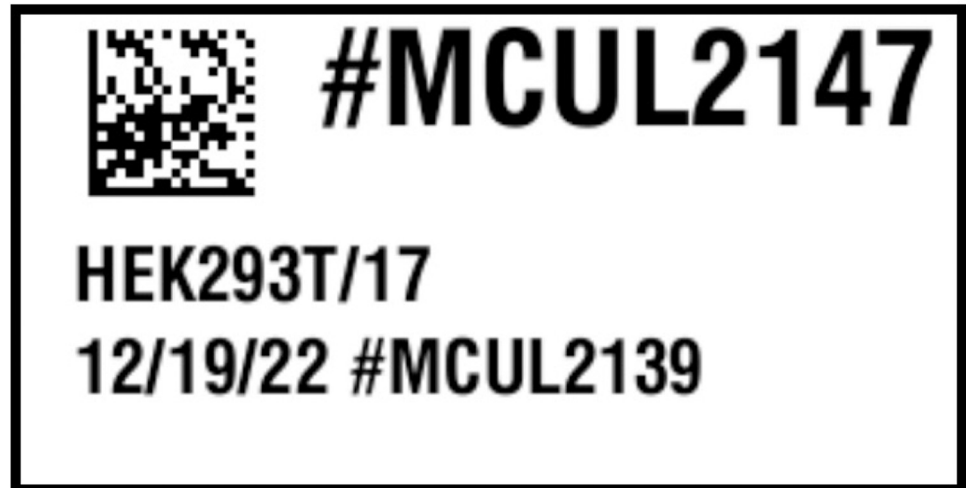


Fig 2. Sample label. The first line includes a unique computer-readable barcode as well as a human-readable computer-generated sample identification number. The second and third lines include a description of the sample content, the date, and the identity of the inoculum.

<https://doi.org/10.1371/journal.pcbi.1011652.g002>

read without the use of any equipment or while doing manual operations (i.e., handling samples).

To make an identifier readable by humans, it is best to keep the ID short and use their structure to provide contextual information. For example, the use of a prefix may help interpret the ID. For example, the ID #CHEM1234 would naturally be interpreted as a chemical or #MCUL2147 as a mammalian culture (Fig 2).

Since these identifiers do not need to map to a unique database entry, human-readable IDs do not have the same uniqueness requirements as computer-readable IDs. For example, it may be acceptable to allow 2 labs using the same software environment to use the same human-readable ID, because this ID will only be seen in the context of a single lab. The software system should maintain the integrity of the relationships between the human-readable ID and the computer-readable ID by preventing users from editing these identifiers.

Rule 5: Label everything

Print labels to identify supplies, equipment, samples, storage locations, and any other physical objects used in your lab. Many labs are extensively relying on handwritten markings that create numerous problems [17]. A limited amount of information can be written on small sample containers, and like manually generated sample names, handwritten labels can be difficult to read or interpret.

Some labels are self-contained. For example, shipping labels include all the information necessary to deliver a package. However, in a laboratory environment, a sample label must not only identify a physical sample but also establish a connection to a record describing the sample and the data associated with it (Fig 2).

Content of a label

Only 2 pieces of information are necessary on a label: a computer-readable Sample ID printed as a barcode and a human-readable Sample ID to make it easier for the researcher to work with the sample. If there is enough space to print more information on the label, your research needs should inform your label design. Ensure you have sufficient space to meet regulatory

labeling requirements (e.g., biosafety requirements, hazards) and if desired, information such as the sample type, sample creator, date (e.g., of generation or expiration), or information about related samples (e.g., parent/child samples).

Label printing technology

Putting in place a labeling solution requires the integration of several elements, but once configured, proper use of label printing technologies makes it much faster and easier to print labels than to label tubes manually.

There are many types of label printers on the market today and most are compatible with the Zebra Programming Language (ZPL) standard [32]. Labeling laboratory samples can be challenging due to harsh environmental conditions: exposure to liquid nitrogen or other chemicals, as well as ultra-low or high temperatures, will require specialized labels. For labeling plastic items, thermal transfer will print the most durable labels, especially if used with resin ribbon instead of wax, while inkjet printers can print durable labels for use on paper [33–35]. Furthermore, laboratory samples can be generated in a broad range of sizes, so labels should be adapted to the size of the object they are attached to. A high-resolution printer (300 dpi or greater) will make it possible to print small labels that will be easy to read by humans and scanners. Finally, select permanent or removable labels based on the application. Reusable items should be labeled with removable labels, whereas single-use containers are best labeled with permanent labels.

Label printing software applications can take data from a database or a spreadsheet and map different columns to the fields of label templates, helping to standardize your workflows (Rule 8). They also support different formats and barcode standards. Of course, the label printing software needs to be compatible with the label printer. When selecting a barcode scanner, consider whether it supports the barcode standards that will be used in your label, as well as the size and shape of the barcodes it can scan. Inexpensive barcode scanners will have difficulty reading small barcodes printed on curved tubes with limited background, whereas professional scanners with high performance imagers will effectively scan more challenging labels. When used, barcode scanners transmit a unique series of characters to the computer. How these characters are then used depends on the software application in which the barcode is read. Some applications will simply capture the content of the barcode. Other applications will process barcoded data in real-time to retrieve the content of the corresponding records.

Rule 6: Manage your data proactively

Many funding agencies now require investigators to include a data management and sharing plan with their research proposals [36,37], and journals have data-sharing policies that authors need to uphold [38]. However, the way many authors share their data indicates a poor understanding of data management [39,40]. Data should not be managed only when publishing the results of a project, they should be managed before the data collection starts [41]. Properly managed data will guide project execution by facilitating analysis as data gets collected (Fig 1). Projects that do not organize their data will face difficulties during analysis, or worse, a loss of critical information that will negatively impact progress.

Use databases to organize your data

It can be tempting to only track data files through notebook entries or dump them in a shared drive (more in Rule 9). That simple data management strategy makes it very difficult to query data that may be spread across multiple files or runs, especially because a lot of contextual information must be captured in file names and directory structures using conventions that

Table 1. Comparison of data management frameworks.

	Relational Databases	NoSQL Databases	Data Lakes	Data Warehouses
Primary Application	Querying and analyzing structured data organized in tables (i.e., columns and rows)	Storing and querying unstructured and semi-structured data; type depends on data model	Storing raw/pre-processed structured and unstructured data	Storing and analyzing large-scale structured data (often SQL-like)
Advantages	<ul style="list-style-type: none"> Widely used Consistent structure Standardized structured query language (SQL) 	<ul style="list-style-type: none"> Data schema is flexible Complementary to relational databases 	<ul style="list-style-type: none"> Simple data ingestion Facilitates analysis of heterogeneous data 	<ul style="list-style-type: none"> Higher storage capabilities Optimized for analysis
Limitations	<ul style="list-style-type: none"> Scalability (performance decreases as data volume increases) Data schema is fixed 	<ul style="list-style-type: none"> Often optimized for certain use cases No standard query language Analysis capabilities depend on technology 	<ul style="list-style-type: none"> Metadata management is critical for data access and interrogation 	<ul style="list-style-type: none"> Data must be processed for ingestion Typically process-focused (potential for lower scope)
Example technologies	<ul style="list-style-type: none"> MySQL Oracle Microsoft SQL server IBM Db2 PostgreSQL Amazon Relational Database Service (RDS) 	Document-oriented: <ul style="list-style-type: none"> MongoDB Key-value pair: <ul style="list-style-type: none"> DynamoDB Column-oriented: <ul style="list-style-type: none"> Apache Cassandra Google Bigtable Graph-based: <ul style="list-style-type: none"> Neo4j 	<ul style="list-style-type: none"> Amazon S3 Microsoft Azure Data Lake Apache Hadoop Google Cloud Storage *Databricks 	<ul style="list-style-type: none"> *Snowflake Data Cloud Amazon Redshift Google BigQuery Apache Hive IBM Db2 Warehouse

*These technologies have been dubbed “lakehouses” that integrate components of data lakes and data warehouses.

<https://doi.org/10.1371/journal.pcbi.1011652.t001>

are difficult to enforce. Today, most data are produced by computer-controlled instruments that export tabular data (i.e., rows and columns) that can easily be imported into relational databases. Data stored in relational databases (e.g., MySQL) are typically explored using standard query language (SQL) and can be easily analyzed using a variety of statistical methods (Table 1). There are also no-code and low-code options, such as the Open Science Framework (<https://osf.io/>) [42], AirTable, and ClickUp, which can also be used to track lab processes, develop standardized workflows, manage teams, etc.

In the age of big data applications enabled by cloud computing infrastructures, there are more ways than ever to organize data. Today, NoSQL (not only SQL) databases [43–45], data lakes [46–48], and data warehouses [49,50] provide additional avenues to manage complex sets of data that may be difficult to manage in relational databases (Table 1). All these data management frameworks make it possible to query and analyze data, depending on the size, type, and structure of your data as well as your analysis goals. NoSQL databases can be used to store and query data that is unstructured or otherwise not compatible with relational databases. Different NoSQL databases implement different data models to choose from depending on your needs (Table 1). Data lakes are primarily used for storing large-scale data with any structure. It is easy to input data into a data lake, but metadata management is critical for organizing, accessing, and interrogating the data. Data warehouses are best suited for storing and analyzing large-scale structured data. They are often SQL-like and are sometimes optimized for specific analytical workflows. These technologies are constantly evolving and the overlap between them is growing as captured in the idea of “lakehouses” such as Databricks and Snowflake Data Cloud (<https://www.snowflake.com/en/>) (Table 1).

When choosing a data management system, labs must consider the trade-off between the cost of the service and the accessibility of the data (i.e., storage in a data lake may be cheaper than in a data warehouse, but retrieving/accessing the data may be more time-consuming or costly) [51]. Many companies offer application programming interfaces (API) to connect their instruments and/or software to databases. In addition, new domain-specific databases

continue to be developed [52]. If necessary, it is also possible to develop your own databases for particular instruments or file types [53]. Nevertheless, when uploading your data to a database, it is recommended to import them as interoperable nonproprietary file types (e.g., .csv instead of .xls for tabular data; .gb (GenBank flat file <https://www.ncbi.nlm.nih.gov/genbank/>) instead of .clc (Qiagen CLC Sequence Viewer format [54]) for gene annotation data; see Rule 4 of [51] for more), so that the data can be accessed if a software is unavailable for any reason and to facilitate data sharing using tools such as git (Rule 10) [14,24].

Link data to protocols

One of the benefits of data organization is the possibility of capturing critical metadata describing how the data were produced. Many labs have spent years refining protocols to be used in different experiments. Many of these protocols have minor variations that can significantly alter the outcome of an experiment. If not properly organized, this can cause major reproducibility issues and can be another uncontrolled source of technical variation. By linking protocol versions to the associated data that they produced (ideally all the samples generated throughout the experiment), it is possible to use this metadata to inform data reproducibility and analysis efforts.

Capture context in notebook entries

Organizing data in databases and capturing essential metadata describing the data production process can greatly simplify the process of documenting research projects in laboratory notebooks [55]. Instead of needing to include copies of the protocols and the raw data produced by the experiment, the notebook entry can focus on the context, purpose, and results of the experiment. In the case of electronic lab notebooks (ELNs; e.g., [SciNote](#), [LabArchives](#), and [eLabJournal](#)), entries can benefit from providing links to previous notebook entries, the experimental and analytical protocols used, and the datasets produced by the workflows. ELNs also bring additional benefits like portability, standardized templates, and improved reproducibility. Finally, notebook entries should include the interpretation of the data as well as a conclusion pointing to the next experiment. The presence of this rich metadata and detailed provenance is critical to ensuring the FAIR principles are being met and your experiments are reproducible [24].

Rule 7: Separate parameters and variables

Not all the data associated with an experiment are the same. Some data are controlled by the operator (i.e., parameters), whereas other data are observed and measured (i.e., variables). It is necessary to establish a clear distinction between set parameters and observed variables to improve reproducibility and analysis.

When parameters are not clearly identified, lab personnel may be tempted to change parameter values every time they perform experiments, which will increase the variability of observations. If, instead, parameter values are clearly identified and defined, then the variance of the observations produced by this set of parameters should be smaller than the variance of the observations produced using different parameter values.

Separating and recording the parameters and variables associated with an experiment makes it possible to build statistical models that compare the observed variables associated with different parameter values [41,56]. It also enables researchers to identify and account for both the underlying biological factors of interest (e.g., strain, treatment) and the technical and random (noise) sources of variation (e.g., batch effects) in an experiment [56].

Utilizing metadata files is a convenient way of reducing variability caused by parameter value changes. A metadata file should include all the parameters needed to perform the same

experiment with the same equipment. In an experimental workflow, pairing a metadata file with the quantified dataset is fundamental to reproducing the same experiment later [51,55,57]. Additionally, metadata files allow the user to assess whether multiple experiments were performed using the same parameters.

Rule 8: Standardize your workflows

Track your parameters from beginning to end

Experimental parameters have a direct influence on observations. However, some factors may have indirect effects on observations or affect observations further downstream in a pipeline. For example, the parameters of a DNA purification process may indirectly influence the quality of sequencing data derived from the extracted DNA.

To uncover such indirect effects, it is necessary to capture the sequence of operations in workflows. For the above example, this would include the DNA extraction, preparation of the sequencing library, and the sequencing run itself. When dealing with such workflows, it is not possible to use a single Sample ID as the key connecting different datasets as in Rule 4. The workflow involves multiple samples (i.e., the biological sample or tissue, the extracted DNA, the sequencing library) that each have their own identifier. Comprehensive inventory and data management systems will allow you to track the sample lineage and flows of data produced at different stages of an experimental process.

Recording experimental parameters and workflows is especially critical when performing new experiments, since they are likely to change over time. As they are finalized, this information can be used to develop both standardized templates for documenting your workflow, as well as metrics for defining the success of each experiment, which can help you to optimize your experimental design and data collection efforts (Fig 1).

Document your data processing pipeline

After the experimental data are collected, it is important to document the different steps used to process and analyze the data, such as if normalization was applied to the data, or if extreme values were not considered in the analyses. The use of ELNs and LIMS can facilitate standardized documentation: creating templates for experimental and analysis protocols can ensure that all the necessary information is collected, thereby improving reproducibility and publication efforts [55,58].

Similarly, thorough source-code documentation is necessary to disseminate your data and ensure that other groups can reproduce your analyses. There are many resources on good coding and software engineering practices [14–16,59], so we only touch on a few important points. Developing a “computational narrative” by writing comments alongside your code or using interfaces that allow for markdown (e.g., Jupyter notebooks, R Markdown) can make code more understandable [60–62]. Additionally, using syntax conventions and giving meaningful names to your code variables increases readability (i.e., use `average_mass = 10` instead of `am = 10`). Furthermore, documenting the libraries or software used and their versions is necessary to achieve reproducibility. Finally, implementing a version control system, such as git, protects the provenance of your work and enables collaboration [63].

Rule 9: Avoid data silos

Depending on your workflows, you may collect information from different instruments or use several databases to store and interact with different types of data. Care must be taken to prevent any of these databases from becoming data silos: segregated groups of data that restrict

collaboration and make it difficult to capture insights resulting from data obtained by multiple instruments [47,49,64]. Data lakes and data warehouses are good solutions for integrating data silos [47,49,64].

Data silos not only stymie research efforts but also raise significant security issues when the silo is the sole storage location. Keeping your data management plan up-to-date with your current needs and utilizing the right databases for your needs can prevent this issue (Rule 6). Regardless, it is crucial to back up your data in multiple places for when a file is corrupted, a service is unavailable, etc. Optimally, your data should always be saved in 3 different locations: 2 on-site and 1 in the cloud [51]. Of course, precautions should always be taken to ensure the privacy and security of your data online and in the cloud [65,66].

Never delete data

As projects develop and data accumulates, it may be tempting to delete data that no longer seems relevant. Data may also be lost as computers are replaced, research personnel leave, and storage limits are reached. Poorly managed data can be easily lost simply because it is unorganized and difficult to query. However, while data collection remains expensive, data storage continues to get cheaper, so there is little excuse for losing or deleting data today. The exception may be intermediary data that is generated by reproducible data processing pipelines, which can be easily regenerated if and when necessary. Most data files can also be compressed to combat limitations on storage capacity.

Properly organized data is a treasure trove of information waiting to be discovered. By using computer-generated sample IDs (Rule 4) and data lakes/warehouses (Rule 6) to link data collected on different instruments, it is possible to extract and synthesize more information than originally intended in the project design. Data produced by different projects using common workflows (Rule 8) can be analyzed to improve workflow performance. Data from failed experiments can be used to troubleshoot a problem affecting multiple projects.

Rule 10: Place your lab in context

Once you have developed a common culture (Rule 1), inventoried your laboratory (Rules 2 and 3), labeled your samples and equipment with computer-generated IDs (Rules 4 and 5), standardized your parameters and workflows (Rules 7 and 8), and backed up your data in several databases (Rules 6 and 9), what comes next?

Track processes occurring outside the lab

Laboratory operations and the data they produce are increasingly dependent on operations that take place outside of the lab. For example, the results of a PCR assay will be affected by the primer design algorithm and the values of its parameters. They will also be affected by the quality of the primers manufactured by a specialized service provider. Even though the primer design and primer synthesis are not taking place in the lab, they are an integral part of the process of generating PCR data. They should therefore be captured in data flows (Rule 8). Furthermore, the software and computational steps used to design experiments and analyze data they produce must also be properly recorded, to identify as many factors that may affect the data produced in the lab as possible.

Increase the accessibility of your work

There are several ways to place your lab in the greater scientific context and increase reproducibility. As discussed, using standardized, non-proprietary file types can increase ease of access

within a lab and across groups [14,51]. You may also choose to make your data and source code public in an online repository to comply with journal requirements, increase transparency, or allow access to your data by other groups [67]. In addition, data exchange standards, such as the Synthetic Biology Open Language [68,69], increase the accessibility and reproducibility of your work.

Practice makes perfect

Whereas traditional data management methods can restrict your analyses to limited subsets of data, centralized information management systems (encompassing relational and NoSQL databases, metadata, sample tracking, etc.) facilitate the analysis of previously disparate datasets. Given the increasing availability and decreasing cost of information management systems, it is now possible for labs to produce, document, and track a seemingly endless amount of samples and data, and use these to inform their research directions in previously impossible ways. When establishing your LIMS, or incorporating new experiments, it is better to capture more data than less. As you standardize your workflows (Rule 8), you should be able to establish clear metrics defining the success of an experiment and to scale the amount of the data you collect as needed.

While there are plenty of existing ELN and LIMS services to choose from (see Rule 6 and Rule 2, respectively), none are a turnkey solution (S1 Supporting information). All data management systems require configuration and optimization for an individual lab. Each service has its own benefits and limitations your group must weigh. Coupled with the need to store your data with multiple backup options, thoughtful management practices are necessary to make any of these technologies work for your lab. The 10 rules discussed here should provide both a starting place and continued resource in the development of your lab information management system. Remember that developing a LIMS is not a one-time event; all lab members must contribute to the maintenance of the LIMS and document their supplies, samples, and experiments in a timely manner. Although it might be an overwhelming process to begin with, careful data management will quickly benefit the data, users, and lab through saved time, standardized practices, and more powerful insights [17,18].

Conclusions

Imparting a strong organizational structure for your lab information can ultimately save you both time and money if properly maintained. We present these 10 rules to help you build a strong foundation in managing your lab information so that you may avoid the costly and frustrating mistakes we have made over the years. By implementing these 10 rules, you should see some immediate benefits of your newfound structure, perhaps in the form of extra fridge space or fewer delays waiting for a reagent you did not realize was exhausted. In time, you will gain deep insights into your workflows and more easily analyze and report your data. The goal of these rules is also to spur conversation about lab management systems both between and within labs as there is no one-size-fits-all solution for lab management. While these rules provide a great starting point, the topic of how to manage lab information is something that must be a constant dialogue. The lab needs to discuss what is working and what is not working to assess and adjust the system to meet the needs of the lab. This dialogue must also be extended to all new members of the lab as many of these organizational steps may not be intuitive. It is critical to train new members extensively and to ensure that they are integrated into the lab's common culture or else you risk falling back into bad practices. If properly trained, lab members will propagate and benefit from the organizational structure of the lab.

Author Contributions

Software: Samuel J. Peccoud.

Validation: Shuyi Ma, Dequina A. Nicholas.

Writing – original draft: Casey-Tyler Berezin.

Writing – review & editing: Luis U. Aguilera, Sonja Billerbeck, Philip E. Bourne, Douglas Densmore, Paul Freemont, Thomas E. Gorochowski, Sarah I. Hernandez, Nathan J. Hillson, Connor R. King, Michael Köpke, Katie M. Miller, Tae Seok Moon, Jason H. Moore, Brian Munsky, Chris J. Myers, Dequina A. Nicholas, Samuel J. Peccoud, Wen Zhou.

References

1. Noble D. The rise of computational biology. *Nat Rev Mol Cell Biol.* 2002; 3(6):459–463. <https://doi.org/10.1038/nrm810> PMID: 12042768
2. Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, et al. Current progress and open challenges for applying deep learning across the biosciences. *Nat Commun.* 2022; 13(1). <https://doi.org/10.1038/s41467-022-29268-7> PMID: 35365602
3. Chuang HY, Hofree M, Ideker T. A Decade of Systems Biology. *Annu Rev Cell Dev Biol.* 2010; 26:721–744. <https://doi.org/10.1146/annurev-cellbio-100109-104122> WOS:000284856700028. PMID: 20604711
4. Chen C, Le H, Goudar CT. Integration of systems biology in cell line and process development for biopharmaceutical manufacturing. *Biochem Eng J.* 2016; 107:11–17. <https://doi.org/10.1016/j.bej.2015.11.013> WOS:000369451800002.
5. Yue R, Dutta A. Computational systems biology in disease modeling and control, review and perspectives. *NPJ Syst Biol Appl.* 2022; 8(1). <https://doi.org/10.1038/s41540-022-00247-4> PMID: 36192551
6. Markowitz F. All biology is computational biology. *PLoS Biol.* 2017; 15(3). ARTN e2002050 <https://doi.org/10.1371/journal.pbio.2002050> WOS:000397909600024. PMID: 28278152
7. Ghosh S, Matsuoka Y, Asai Y, Hsin KY, Kitano H. Software for systems biology: from tools to integrated platforms. *Nat Rev Genet.* 2011; 12(12):821–832. <https://doi.org/10.1038/nrg3096> WOS:000297252500009. PMID: 22048662
8. El Karoui M, Hoyos-Flight M, Fletcher L. Future Trends in Synthetic Biology—A Report. *Front Bioeng Biotech.* 2019; 7. <https://doi.org/10.3389/fbioe.2019.00175> WOS:000479268000001. PMID: 31448268
9. Osborne JM, Bernabeu MO, Bruna M, Calderhead B, Cooper J, Dalchau N, et al. Ten Simple Rules for Effective Computational Research. *PLoS Comput Biol.* 2014; 10(3):e1003506. <https://doi.org/10.1371/journal.pcbi.1003506> PMID: 24675742
10. Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput Biol.* 2013; 9(10):e1003285. <https://doi.org/10.1371/journal.pcbi.1003285> PMID: 24204232
11. Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol.* 2014; 10(4):e1003542. <https://doi.org/10.1371/journal.pcbi.1003542> PMID: 24763340
12. Kazic T. Ten Simple Rules for Experiments' Provenance. *PLoS Comput Biol.* 2015; 11(10):e1004384. Epub 20151020. <https://doi.org/10.1371/journal.pcbi.1004384> PMID: 26485673; PubMed Central PMCID: PMC4619002.
13. Zook M, Barocas S, Boyd D, Crawford K, Keller E, Gangadharan SP, et al. Ten simple rules for responsible big data research. *PLoS Comput Biol.* 2017; 13(3):e1005399. <https://doi.org/10.1371/journal.pcbi.1005399> PMID: 28358831
14. List M, Ebert P, Albrecht F. Ten Simple Rules for Developing Usable Software in Computational Biology. *PLoS Comput Biol.* 2017; 13(1):e1005265. <https://doi.org/10.1371/journal.pcbi.1005265> PMID: 28056032
15. Plić A, Procter JB. Ten Simple Rules for the Open Development of Scientific Software. *PLoS Comput Biol.* 2012; 8(12):e1002802. <https://doi.org/10.1371/journal.pcbi.1002802> PMID: 23236269
16. Taschuk M, Wilson G. Ten simple rules for making research software more robust. *PLoS Comput Biol.* 2017; 13(4):e1005412. <https://doi.org/10.1371/journal.pcbi.1005412> PMID: 28407023

17. Myneni S, Patel VL. Organization of biomedical data for collaborative scientific research: A research information management system. *Int J Inf Manag*. 2010; 30(3):256–264. <https://doi.org/10.1016/j.ijinfomgt.2009.09.005> PMID: 20543892
18. Prasad PJ, Bodhe GL. Trends in laboratory information management system. *Chemom Intell Lab Syst*. 2012; 118:187–92. <https://doi.org/10.1016/j.chemolab.2012.07.001>
19. Knapp B, Bardenet R, Bernabeu MO, Bordas R, Bruna M, Calderhead B, et al. Ten Simple Rules for a Successful Cross-Disciplinary Collaboration. *PLoS Comput Biol*. 2015; 11(4):e1004214. <https://doi.org/10.1371/journal.pcbi.1004214> PMID: 25928184
20. Macleod M. What makes interdisciplinarity difficult? Some consequences of domain specificity in interdisciplinary practice. *Synthese*. 2018; 195(2):697–720. <https://doi.org/10.1007/s11229-016-1236-4>
21. Vicens Q, Bourne PE. Ten Simple Rules for a Successful Collaboration. *PLoS Comput Biol*. 2007; 3(3):e44. <https://doi.org/10.1371/journal.pcbi.0030044> PMID: 17397252
22. Andreev A, Komatsu V, Almiron P, Rose K, Hughes A, Lee MY. Welcome to the lab. *elife*. 2022;11. <https://doi.org/10.7554/eLife.79627> PMID: 35503004
23. Rasmussen L, Maddox CB, Harten B, White EL. A Successful LIMS Implementation: Case Study at Southern Research Institute. *J Assoc Lab Autom*. 2016; 12(6):384–390. <https://doi.org/10.1016/j.jala.2007.08.002>
24. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016; 3(1):160018. <https://doi.org/10.1038/sdata.2016.18> PMID: 26978244
25. Perkel JM. Lab-inventory management: Time to take stock. *Nature*. 2015; 524(7563):125–126. <https://doi.org/10.1038/524125a> PMID: 26245582
26. Foster BL. The Chemical Inventory Management System in academia. *Chem Health Saf*. 2005; 12(5):21–5. <https://doi.org/10.1016/j.chs.2005.01.019>
27. Sivakami N. Comparative study of barcode, QR-code and RFID system in library environment. *International Journal of Academic Research. Libr Inf Sci*. 2018; 1(1).
28. Tiwari S, editor. An Introduction to QR Code Technology. 2016 International Conference on Information Technology (ICIT); 2016. 22–24.
29. Copp AJ, Kennedy TA, Muehlbauer JD. Barcodes Are a Useful Tool for Labeling and Tracking Ecological Samples. *Bull Ecol Soc Am*. 2014; 95(3):293–300. <https://doi.org/10.1890/0012-9623-95.3.293>
30. Shukran MAMI Ishak MS, Abdullah MN. Enhancing Chemical Inventory Management in Laboratory through a Mobile-Based QR Code Tag. *IOP Conference Series: Materials Science and Engineering*. 2017; 226(012093). <https://doi.org/10.1088/1757-899x/226>
31. Wahab MHA, Kadir HA, Tukiran Z, Tomari MR, Mutalib AA, Mohsin MFM, et al., editors. Web-based laboratory equipment monitoring system using RFID2010: IEEE.
32. Zebra Developers. Zebra Programming Language (ZPL) [09/30/2023]. Available from: <https://developer.zebra.com/products/printers/zpl>.
33. Beiner G, editor Labels for Eternity: Testing Printed Labels for use in Wet Collections. *Collection Forum*; 2020: Soc. for the Pres. of Natural History Collections.
34. CableOrganizer. Thermal Transfer Label Printers and How They Work [09/30/2023].
35. Technicode Inc. How to Select the Right Thermal Transfer Ribbon Type 2022 [09/30/2023]. Available from: <https://technicodelabels.com/blog/how-to-select-thermal-transfer-ribbon-type/>.
36. Gonzales S, Carson MB, Holmes K. Ten simple rules for maximizing the recommendations of the NIH data management and sharing plan. *PLoS Comput Biol*. 2022; 18(8):e1010397. Epub 20220803. <https://doi.org/10.1371/journal.pcbi.1010397> PMID: 35921268; PubMed Central PMCID: PMC9348704.
37. Michener WK. Ten Simple Rules for Creating a Good Data Management Plan. *PLoS Comput Biol*. 2015; 11(10):e1004525. <https://doi.org/10.1371/journal.pcbi.1004525> PMID: 26492633
38. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*. 2016; 3:160018. Epub 20160315. <https://doi.org/10.1038/sdata.2016.18> PMID: 26978244; PubMed Central PMCID: PMC4792175.
39. Peccoud J. Data Sharing Policies: Share Well and You Shall be Rewarded. *Synth Biol*. 2021. <https://doi.org/10.1093/synbio/ysab028> PMID: 34604538
40. Christensen G, Dafoe A, Miguel E, Moore DA, Rose AK. A study of the impact of data sharing on article citations using journal policies as a natural experiment. *PLoS ONE*. 2019; 14(12):e0225883. Epub 2019/12/19. <https://doi.org/10.1371/journal.pone.0225883> PMID: 31851689; PubMed Central PMCID: PMC6919593.

41. Kass RE, Caffo BS, Davidian M, Meng X-L, Yu B, Reid N. Ten Simple Rules for Effective Statistical Practice. *PLoS Comput Biol*. 2016; 12(6):e1004961. <https://doi.org/10.1371/journal.pcbi.1004961> PMID: 27281180
42. Foster ED, Deardorff A. Open Science Framework (OSF). *J Med Libr Assoc*. 2017; 105(2). <https://doi.org/10.5195/jmla.2017.88>
43. Nayak AP, Poriya A, Poojary D. Type of NOSQL Databases and its Comparison with Relational Databases. *Int J Appl*. 2013; 5(4):16–19.
44. Pokorný J. Integration of Relational and NoSQL Databases. *Vietnam J Comput Sci*. 2019; 06(04):389–405. <https://doi.org/10.1142/s2196888819500210>
45. Sahatqija K, Ajdari J, Zenuni X, Raufi B, Ismaili F, editors. Comparison between relational and NOSQL databases. 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). 2018 21–25 May.
46. Nargesian F, Zhu E, Miller RJ, Pu KQ, Arocena PC. Data lake management. *Proc VLDB Endow*. 2019; 12(12):1986–1989. <https://doi.org/10.14778/3352063.3352116>
47. Sawadogo P, Darmont J. On data lake architectures and metadata management. *J Intell Inf Syst*. 2021; 56(1):97–120. <https://doi.org/10.1007/s10844-020-00608-7>
48. Giebler C, Gröger C, Hoos E, Schwarz H, Mitschang B. Leveraging the Data Lake: Current State and Challenges. Springer International Publishing; 2019. p. 179–88.
49. Nambiar A, Mundra D. An Overview of Data Warehouse and Data Lake in Modern Enterprise Data Management. *Big Data Cogn Comput*. 2022; 6(4):132. <https://doi.org/10.3390/bdcc6040132>
50. Rudniy A. Data Warehouse Design for Big Data in Academia. *Comput Mater Contin*. 2022; 71(1):979–992. <https://doi.org/10.32604/cmc.2022.016676>
51. Hart EM, Barmby P, Lebauer D, Michonneau F, Mount S, Mulrooney P, et al. Ten Simple Rules for Digital Data Storage. *PLoS Comput Biol*. 2016; 12(10):e1005097. <https://doi.org/10.1371/journal.pcbi.1005097> PMID: 27764088
52. Rigden DJ, Fernández XM. The 2022 Nucleic Acids Research database issue and the online molecular biology database collection. *Nucleic Acids Res*. 2022; 50(D1):D1–D10. <https://doi.org/10.1093/nar/gkab1195> PMID: 34986604
53. Helmy M, Crits-Christoph A, Bader GD. Ten Simple Rules for Developing Public Biological Databases. *PLoS Comput Biol*. 2016; 12(11):e1005128. <https://doi.org/10.1371/journal.pcbi.1005128> PMID: 27832061
54. Qiagen. The CLC format 2018 [09/29/2023]. Available from: https://resources.qiagenbioinformatics.com/manuals/clcsequenceviewer/current/index.php?manual=Introduction_CLC_Sequence_Viewer.html.
55. Nussbeck SY, Weil P, Menzel J, Marzec B, Lorberg K, Schwappach B. The laboratory notebook in the 21st century. *EMBO Rep*. 2014; 15(6):631–634. <https://doi.org/10.15252/embr.201338358> PMID: 24833749
56. Altman N, Krzywinski M. Sources of variation. *Nat Methods*. 2015; 12(1):5–6. <https://doi.org/10.1038/nmeth.3224> PMID: 25699313
57. White E, Baldrige E, Brym Z, Locey K, McGlenn D, Supp S. Nine simple ways to make it easier to (re) use your data. *Ideas in Ecology and Evolution*. 2013; 6(2). <https://doi.org/10.4033/iee.2013.6b.6.f>
58. Dunie M. The importance of research data management: The value of electronic laboratory notebooks in the management of data integrity and data availability. *Inf Serv Use*. 2017; 37:355–359. <https://doi.org/10.3233/ISU-170843>
59. Lee BD. Ten simple rules for documenting scientific software. *PLoS Comput Biol*. 2018; 14(12):e1006561. <https://doi.org/10.1371/journal.pcbi.1006561> PMID: 30571677
60. Baumer B, Udwin D. R Markdown. *Wiley Interdiscip Rev Comput Stat*. 2015; 7(3):167–177. <https://doi.org/10.1002/wics.1348>
61. Kluyver TR-K, Perez B, Granger F, Bussonnier B, Frederic M, Kelley J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. UC Berkeley. 2016. <https://doi.org/10.3233/978-1-61499-649-1-87>
62. Rule A, Birmingham A, Zuniga C, Altintas I, Huang S-C, Knight R, et al. Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Comput Biol*. 2019; 15(7):e1007007. <https://doi.org/10.1371/journal.pcbi.1007007> PMID: 31344036
63. Perez-Riverol Y, Gatto L, Wang R, Sachsenberg T, Uszkoreit J, Leprevost FDV, et al. Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Comput Biol*. 2016; 12(7):e1004947. <https://doi.org/10.1371/journal.pcbi.1004947> PMID: 27415786

64. Bridging PJ, Integration DSUBD. *International Journal of Database Management Systems*. 2019; 11(3):01–06. <https://doi.org/10.5121/ijdms.2019.11301>
65. Abadi D. *Data Management in the Cloud: Limitations and Opportunities*. *IEEE Data Eng Bull*. 2009; 32:3–12.
66. Tabrizchi H, Kuchaki RM. A survey on security challenges in cloud computing: issues, threats, and solutions. *J Supercomput*. 2020; 76(12):9493–9532. <https://doi.org/10.1007/s11227-020-03213-1>
67. Boland MR, Karczewski KJ, Tatonetti NP. Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. *PLoS Comput Biol*. 2017; 13(1):e1005278. <https://doi.org/10.1371/journal.pcbi.1005278> PMID: [28103227](https://pubmed.ncbi.nlm.nih.gov/28103227/)
68. Baig H, Fontanarrosa P, Kulkarni V, McLaughlin JA, Vaidyanathan P, Bartley B, et al. Synthetic biology open language (SBOL) version 3.0.0. *J Integr Bioinform*. 2020; 17(2–3). Epub 20200625. <https://doi.org/10.1515/jib-2020-0017> PMID: [32589605](https://pubmed.ncbi.nlm.nih.gov/32589605/); PubMed Central PMCID: [PMC7756618](https://pubmed.ncbi.nlm.nih.gov/PMC7756618/).
69. Galdzicki M, Clancy KP, Oberortner E, Pocock M, Quinn JY, Rodriguez CA, et al. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat Biotechnol*. 2014; 32(6):545–550. <https://doi.org/10.1038/nbt.2891> PMID: [24911500](https://pubmed.ncbi.nlm.nih.gov/24911500/).