





## OPINION ARTICLE

# REVISED Ten recommendations for organising bioimaging data for archival [version 2; peer review: 3 approved, 1 approved with reservations]

Paul K. Korir , Andrii Iudin, Sriram Somasundharam, Simone Weyand, Osman Salih, Matthew Hartley, Ugis Sarkans, Ardan Patwardhan, Gerard J. Kleywegt 

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, UK

**V2** First published: 23 Oct 2023, 12(ELIXIR):1391  
<https://doi.org/10.12688/f1000research.129720.1>













Latest published: 27 Feb 2024, 12(ELIXIR):1391  
<https://doi.org/10.12688/f1000research.129720.2>





## Abstract

Organised data is easy to use but the rapid developments in the field of bioimaging, with improvements in instrumentation, detectors, software and experimental techniques, have resulted in an explosion of the volumes of data being generated, making well-organised data an elusive goal. This guide offers a handful of recommendations for bioimage depositors, analysts and microscope and software developers, whose implementation would contribute towards better organised data in preparation for archival. Based on our experience archiving large image datasets in EMPIAR, the BioImage Archive and BioStudies, we propose a number of strategies that we believe would improve the usability (clarity, orderliness, learnability, navigability, self-documentation, coherence and consistency of identifiers, accessibility, succinctness) of future data depositions more useful to the bioimaging community (data authors and analysts, researchers, clinicians, funders, collaborators, industry partners, hardware/software producers, journals, archive developers as well as interested but non-specialist users of bioimaging data). The recommendations that may also find use in other data-intensive disciplines. To facilitate the process of analysing data organisation, we present bandbox, a Python package that provides users with an assessment of their data by flagging potential issues, such as redundant directories or invalid characters in file or folder names, that should be addressed before archival. We offer these recommendations as a starting point and hope to engender more substantial conversations across and between the various data-rich

## Open Peer Review

Approval Status 

	1	2	3	4
<b>version 2</b> (revision) 27 Feb 2024	 <a href="#">view</a>	 <a href="#">view</a>	 <a href="#">view</a>	 <a href="#">view</a>
				
<b>version 1</b> 23 Oct 2023	 <a href="#">view</a>	 <a href="#">view</a>	 <a href="#">view</a>	 <a href="#">view</a>

1. **Sjors Scheres** , Medical Research Council Laboratory of Molecular Biology, Cambridge, UK
  2. **Kenneth H. L. Ho**, The Francis Crick Institute, London, UK
  3. **Sylvia Emmanuelle Le Dévédec** , Universiteit Leiden, Leiden, The Netherlands
  4. **William T. Katz** , Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, USA
- Virginia Scarlett** , Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, USA  
 Howard Hughes Medical Institute's Janelia

communities.

### Keywords

Organising data, public archiving, data deposition, open data, bioimaging, EMPIAR, BioImage Archive, BioStudies

Research Campus, Ashburn, USA

Any reports and responses or comments on the article can be found at the end of the article.



This article is included in the **ELIXIR** gateway.

**Corresponding authors:** Ardan Patwardhan ([ardan@ebi.ac.uk](mailto:ardan@ebi.ac.uk)), Gerard J. Kleywegt ([gerard@ebi.ac.uk](mailto:gerard@ebi.ac.uk))

**Author roles:** **Korir PK:** Conceptualization, Formal Analysis, Methodology, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Iudin A:** Data Curation, Writing – Review & Editing; **Somasundharam S:** Writing – Review & Editing; **Weyand S:** Data Curation, Writing – Review & Editing; **Salih O:** Data Curation, Writing – Review & Editing; **Hartley M:** Writing – Review & Editing; **Sarkans U:** Writing – Review & Editing; **Patwardhan A:** Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Kleywegt GJ:** Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing

**Competing interests:** No competing interests were disclosed.

**Grant information:** This work was supported by UKRI-MRC and UKRI-BBSRC (grants MR/L007835/1 and MR/P019544/1), the Wellcome Trust (grant 221371/Z/20/Z), and EMBL through contributions from its member states.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Copyright:** © 2024 Korir PK *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**How to cite this article:** Korir PK, Iudin A, Somasundharam S *et al.* **Ten recommendations for organising bioimaging data for archival [version 2; peer review: 3 approved, 1 approved with reservations]** F1000Research 2024, 12(ELIXIR):1391 <https://doi.org/10.12688/f1000research.129720.2>

**First published:** 23 Oct 2023, 12(ELIXIR):1391 <https://doi.org/10.12688/f1000research.129720.1>

**REVISED Amendments from Version 1**

In the revision we have addressed most of the constructive comments and suggestions of the reviewers (as indicated in the individual rebuttals).

This has involved:

- Adding clarifications (e.g., on the intended audience and the usefulness of archived data) or examples (e.g., of personal identifiers) where requested
- Making changes to the phrasing to add context or to make it more explicit, less ambiguous or clearer
- Implementing corrections suggested by the reviewers (including adding a few references)
- Making a few changes to the open-source bandbox software, e.g. to limit the amount of output
- Changing the figures to have white backgrounds to improve their readability and to reduce the amount of black ink required for printing
- Applying a number of minor corrections to interpunction, choice of words, etc.
- Acknowledging the reviewers by name
- Adding and correcting a few references

**Any further responses from the reviewers can be found at the end of the article**

**Introduction**

Scientific data archival has a long history of providing publicly accessible storage of experimental data that typically involves manual and automated curation and annotation with appropriate metadata for reuse by others (Whitlock *et al.* 2010; Rausher *et al.* 2010; Berman *et al.* 2014). In the area of bioimaging, global and public resources such as EMPIAR (Iudin *et al.*, 2016, 2023) and the BioImage Archive (Ellenberg *et al.*, 2018; Hartley *et al.*, 2022) provide a valuable service to the life-science community by supporting the archival and reuse of imaging data, often acquired at considerable cost, in line with the aspirations of the FAIR Guiding Principles (Wilkinson *et al.*, 2016). There are numerous advantages and benefits to reusing bioimaging data, including more economical use of limited resources such as instrumentation and highly skilled technical staff. Moreover, specimens may be unique, costly to acquire, or difficult to reproduce, meaning that such data may only be accessible via archives. Archived data can be mined for reanalysis, verification and validation, and for development of new analytical techniques and software tools, such as machine learning model training. Reuse of such data may also lead to improvements in how it is produced, both technologically and methodologically. As practitioners in bioimaging data archiving, it is our experience that handling large datasets presents several data-management challenges, particularly in recent years with the rapidly increasing volumes of bioimaging data (Ellenberg *et al.*, 2018). For example, it took eight years for EMPIAR to archive a total of one petabyte of data, but the second petabyte took only 14 months (Iudin *et al.*, 2023). Bioimaging datasets may comprise numerous and sometimes very large files in a variety of, sometimes proprietary, formats. Individual files may include multiple channels and time points and data and metadata from several specimens. Besides the raw image data, there may also be a need to archive processed data, reconstructed 3D volumes, segmentations, particle stacks and other derived or related data.

There are two related but distinct avenues for organising data: labelling (metadata) and arranging data items (order). Metadata are essential to make the data useful even though metadata standards are difficult to enforce. Therefore, metadata standardisation has received a lot of attention with initiatives such as Bioschemas, an effort to improve findability of datasets via standardised textual annotations, MIAME (Brazma *et al.*, 2001), recommendations for minimal metadata describing a microarray experiment, and the overarching FAIR Guiding Principles (Wilkinson *et al.*, 2016). For bioimaging, REMBI (Sarkans *et al.*, 2021) provides community-supported recommendations on how to describe all aspects of bioimaging experiments including sample preparation, data processing and analysis. Whereas there are several ongoing efforts towards standardising bioimaging data formats (OME-NGFF (Moore *et al.*, 2021), DVID (Katz & Plaza, 2019), BigDataViewer (Pietzsch *et al.*, 2015), etc.), we know of no efforts towards harmonising how to organise datasets for maximum usefulness for archival in mind. The organisation (order) of data is usually taken for granted and it falls upon refinements of the metadata to bear the burden of meaningfully describing the data. Nevertheless, it is essential to maintain coherence between metadata and order of the data, for example in the naming of entities to facilitate meaningful navigation between the two.

**Motivation**

Good organisation (order) of data improves its usefulness and is the responsibility of the data depositors. Depositors are best placed to present data in a way that adequately captures the experimental design and outcomes. For this reason, the recommendations that are outlined below are primarily targeted at data depositors as they are best positioned to structure the data meaningfully. Additionally, these recommendations are also aimed at developers of software tools that either produce the raw data (acquisition software on microscopes), process and transform it into a useful form, or extract

meaningful domain insights (bioimage analysis). Organising a dataset to minimally convey a structure in line with the actual experimental output can improve its usability while the bulk of meaningful attributes can be expressed in the metadata. The degree of usefulness depends directly on the quality of organisation, and thoughtful consideration of the needs of users (i.e., those who consume the archived data) improves that usefulness. Good organisation also gives a dataset transparency and understandability: users can immediately distinguish the various experimental categories as well as plan how to analyse the data (Petek *et al.*, 2022). Therefore, it helps to have a clear perspective of the various types of users.

In general, we consider three types of users: *intra-domain scientists*, *inter-domain scientists* and *extra-domain scientists* (Datta *et al.*, 2021). (For the purpose of this article, we will refer to any such user of a dataset as a ‘scientist’, interested in extracting some knowledge from the archived data.) Intra-domain scientists are familiar with key attributes of the data and may be able to quickly assess the usefulness of a dataset. An example would be a structural biologist mining an electron cryo-tomogram to extract sub-volumes that have not been previously studied. Inter-domain scientists may want to mine the data for purposes relevant to some other domain. For instance, on the genomics side, using spatial transcriptomics imaging data for fine-grained localization of individual transcripts would be a possible scenario for an inter-domain scientist. Extra-domain scientists are only interested in data for its technical properties, i.e., for some purpose completely unrelated to the original purpose of the data’s collection. A computer scientist, for example, may want to assess the performance of a learning algorithm on fluorescent microscopy images when performing some classification task. It is likely to be a challenge to optimise the organisation of data for all types of users simultaneously. In practice, organising the data to be useful to scientists with the least familiarity with the domain will most likely advance its usefulness for all types of scientists and can thus be a good aspiration.

Organising data results in a hierarchical arrangement of data into files and folders. The visual properties of such an organisation influence the usability of the data. There are several considerations that affect the organisation:

- What are the sets of *symbols* used for naming the files and folders?
- What are the sets of relevant *named entities* these characters describe?
- How does the resulting *hierarchy*, defined using files and folders, capture the relationships between the named entities?

We can refer to the above as *organisational resources*, and it is through their judicious use that the data can become usable. Very long file names, potentially problematic characters, and deep nesting of folders are examples of how injudicious use of these resources can result in unusable data.

A simple example of how this is useful is the way most operating systems apply ordering of a directory’s contents either lexicographically or by other attributes such as date-time stamps. These take advantage of the familiarity that users have with the conventional ordering of these attributes. In non-trivial organisational tasks, we may need to express complex relationships between the entities at hand. For instance, a dataset that consists of the experimental measurements resulting from a sequence of treatments applied on a set of specimens measured at various points in time requires the use of specimen, treatment and time-point identifiers as well as other experimental attributes (data formats, alternative perspectives, transformations of the data such as changes in units, etc.) to be captured in such a way as to preserve the main experimental relationships. In that case, we can expand our set of organisational resources to include file formats in addition to the set of symbols (letters, numerals, punctuation, uppercase and lowercase) used to create the various identifiers. Ideally, we would like to keep repetition to a minimum so that the nature of the experiment can be readily discerned.

The way organisational resources are used affects the usability of the resulting organisation: using too few of them will obscure the meaning of the organisation while using too many will overwhelm potential users. For example, including redundant folders along any part of the hierarchy (folders that contain only a single folder which in turn contains the actual data) makes it tedious to navigate through a dataset. On the other hand, dumping all files into one folder will make it difficult for the end user to distinguish between groups of semantically related files, especially when thousands of files are present. Similarly, naming files and folders by referring to entities inaccessible to their intended users (e.g., local machine names or private accession codes that external users will not have access to or even fathom) consumes precious ‘name space’ without conveying any useful information. Organising data is thus an investment of time and effort with the aim of improving the usefulness of the data.

We can therefore formulate the organisation task as follows: *given a set of related data items associated with an experiment, how may they be organised to best convey their relationships using as few organisational resources as possible while maximising their usability?*

To achieve this, we define the term *facet* to refer to the various attributes germane to the experiment which may be included in the folder and file names. A non-exhaustive list of facets is: specimen names (*organism, tissue, cell type/line*), experimental roles (*treatments vs. controls*), time (*developmental status, date, elapsed time*), processing status (*raw data, by algorithm, procedure*), commonly available experimental equipment (*microscopes, detectors, preparation equipment model names*), replicates, file types (*3D volumes, particle stacks*), names of software used for processing.

This guide attempts to solve the organisation task by providing 10 recommendations that arise from our experience of handling hundreds of large image datasets in the public archives EMPIAR, BioImage Archive and BioStudies (Sarkans *et al.*, 2018). Ideally, we would like to organise potentially numerous and voluminous data to maximise ease of use and hence facilitate the user's ability to:

1. quickly identify the suitability of (subsets of) the data;
2. clearly distinguish between the various facets of the data;
3. quickly verify the usefulness of the data (e.g., thumbnails, previews, summaries, READMEs, LICENCE files);
4. retrieve only relevant subsets of the data.

This guide does not offer any recommendations for a detailed schema to describe experimental and analytical procedures; those may be captured in metadata for the various archives. Neither does it describe how to decide which experimental facets are appropriate (these are part of the experimental design), nor does it attempt to describe how to achieve organisation for automated analysis (we assume that the resulting organisation will be consumed by humans). It also ignores the universe of image formats in use and mainly includes examples from our experience archiving bioimaging data, but we anticipate it may be useful across other imaging disciplines. Good organisation improves data structure and format predictability and may facilitate automated processing. Therefore, our guide is intended to lead towards best practices rather than serve as a framework. Finally, this guide does not aim to achieve standardisation. We believe it is more practical to have a set of best practices and leave it up to the data authors to decide how best to apply them.

We believe that the recommendations outlined here may be of value to two principal groups of users: 1) data depositors, who need to design and prepare their data to improve its usability to the community, and 2) technologists (hardware, software and methods developers), who, by considering these recommendations in their designs, can greatly facilitate good data organisation at the source.

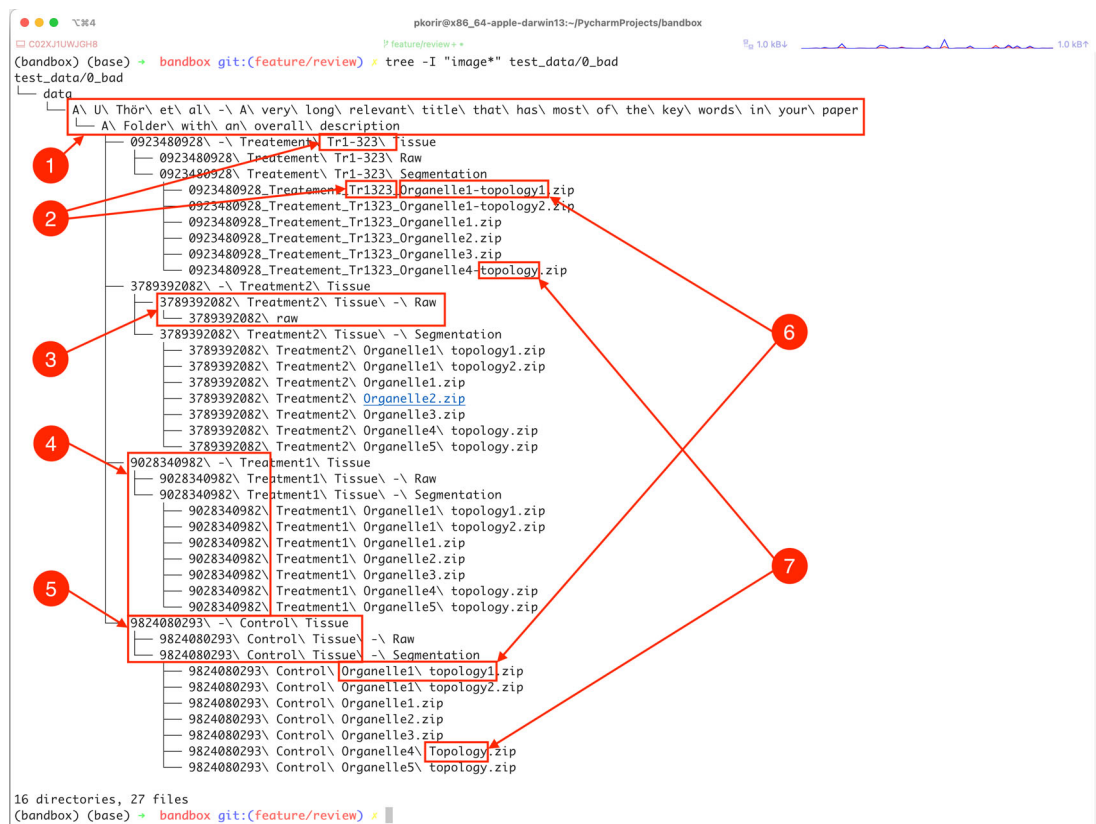
## Recommendations

We will motivate our guide by referring to a fictitious EMPIAR dataset. This dataset has a clear structure, but we propose that it can be further improved following the recommendations in the guide below.

Our goal is to improve the file/folder structure shown in [Figure 1](#) to better convey the relationships between the experimental facets while economising on the organisational resources available. For clarity, we have refrained from listing several thousand uncompressed TIFF files in the folders designated 'Raw'.

The example dataset illustrates several properties of its organisation that undermine the goal of being usable:

- **Verbosity/redundancy** typically manifested in repetition of references which may be resolved using the file hierarchy, such as:
  - Folders containing only a single folder which in turn contains the folder with the actual data. The child folder of 'data' only has the folder 'AUThör et al...' in it that contains the folder 'A folder with an overall description' which has the actual data.



**Figure 1. Illustration of some of the ways in which subtle aspects of data organisation impact its usability.** These include: 1) long file/folder names with spaces, non-ASCII characters (ö) and redundant directories (ASCII – American Standard Code for Information Interchange); 2) obscure names or identifiers, possibly with inconsistent spelling, 3) inconsistency in folder hierarchy, 4) obscurity through meaningless names or identifiers, 5) verbosity in names, 6) subtle differences in spelling (in this case, a hyphen) and 7) inconsistency in typography due to character case and inclusion of different separator characters, e.g., spaces vs hyphens. See text for more details.

- Very long names of files/folders. The full path of the file 'data/A U Thör et al - A very long relevant title that has most of the keywords in your paper/A Folder with an overall description/0923480928 - Treatment Tr1-323 Tissue/0923480928 Treatment Tr1-323 Segmentation/0923480928 Treatment Tr1323 Organelle1-topology1.zip' is '0923480928 Treatment Tr1-323 Segmentation/0923480928 Treatment Tr1323 Organelle1-topology1.zip', which might be outside the limits of legacy software; e.g., IMOD (Mastrorade, 2006) has a limit of 320 characters for input file names.
- Repetition of identifiers along the path. In the previous example, half of the files repeat the identifier '0923480928' that conveys no meaningful information and which, if required at all, should only appear in the appropriate parent folder name.
- **Ambiguity** occurs through incomplete identifiers either due to typos or non-standard characters.
  - Is 'Tr1-323' the same as 'Tr1323'?
  - Use of spaces and non-ASCII characters can make processing the data complicated because of how software may handle path names with spaces. ASCII stands for the American Standard Code for Information Interchange and consists of plain characters used in many languages.
- **Inconsistency** is perhaps the most common issue and is usually the result of manually introduced errors such as changes in spelling, e.g., naming similar folders 'tomo' and 'tomogram' for related files. In the above example we have:

- 'Topology' and 'topology'
- 'Treatment' vs 'Treatement'
- 'Tr1-323' and 'Tr1323'
- Inconsistency may also be observed in folder structure. For example, only one of the treatment folders (the one with '3738932082' in the name) has an extra child folder, breaking the pattern of the others.
- **Obscurity** tends to occur by identifiers with no obvious meaning, e.g., references to external resources such as figure numbers in a related paper, machine identifiers, script names, etc.
  - The numerical identifiers such as '0923480928' have no obvious meaning in the context of the dataset.
  - 'Tr1-323' may be an external reference but its meaning is unclear.

Understandably, in certain cases such obscurity may be useful to keep identifiers which convey additional information. For example, in cryogenic-specimen electron microscopy (cryoEM) pipelines, the dataset may consist of multiple subsets obtained with different open-source software, e.g. particle picking by EMAN2 (Tang *et al.*, 2007), beam-induced motion-correction by MotionCorr (Li *et al.*, 2013), contrast-transfer function (CTF) correction by gCTF (Zhang, 2016), classification by RELION (Scheres, 2012), reconstruction by cryoSPARC (Punjani *et al.*, 2017), etc.

The 10 recommendations we present below are divided into four groups: *planning* (recommendation 1), *structure* (recommendations 2-4), *naming* (recommendations 5-7) and *miscellaneous* (recommendations 8-10). We have provided further guidance within each group for related concepts.

## Planning

**(1) Design before data collection.** Plan beforehand, if possible, how the data will be structured.

- a. If the experimental facets are known prior to data collection, the organisation suggestions that follow below will be easier to apply once and for all; it is harder to reorganise data after collection, especially voluminous data on multiple networked drives or in a cloud resource. At a minimum, consider organising the few top-level directories in terms of the experimental facets prior to archival.
- b. Employ a naming convention within a research group or facility to ensure that data is consistent between data creators. This can even be specified in the microscope's software to include imaging parameters in the file names automatically such as a base name, date and/or time, imaging parameters (e.g., resolution, section size) or even free text, among many others. We invite software vendors/creators that have not already done so to consider taking these recommendations into account.

## Structure

This section contains recommendations to address the hierarchical organisation of files and folders only.

**(2) Top-level folder.** Have one parent folder into which all sub-datasets are located. Such a top-level folder is also a good location to include auxiliary data that apply to the collection such as README or integrity verification files (see recommendation 10), which provide users with the context of the data organisation.

**(3) Filename length, path length and folder depth.**

- a. Impose an upper limit on the length of file and folder names. We propose a working upper limit of 50 characters. Even though modern operating systems have no limitations on the lengths of names, end users will still struggle typing very long names which increases the likelihood of transcription errors. In some cases, older software that is still widely used by the bioimaging community imposes limits on the number of characters for file paths, e.g., IMOD (Mastrorarde, 2006) imposes a file-path length limit of 320 characters. It is useful to bear in mind that, increasingly, users interact with datasets via a web browser, which also has a practical limit (based on the device's memory) on the number of files that can be selected in the browser's select dialog.

- b. Limit the folder depth to a reasonable maximum. As a rule of thumb, three to four directory levels should be adequate for most applications but the fewer the better. This is in line with the ISA framework (Sansone et al., 2008), which organises metadata in three levels (investigation, study, assay). In contrast, both shallow folder depth, with many and varied file types that are difficult to distinguish, and deep nesting of folders make navigation and selection a challenge.
- c. Exclude intermediate levels of folders that do not convey any additional information. For example, consider a dataset having only TIFF files. Including an additional folder called `tiff` in the path `<condition>/tiff/files*.tif` is redundant. By contrast, if the file format is important then `<condition>/<format1>/<files_of_format1>` and `<condition>/<format2>/<files_of_format2>` is meaningful.
- d. Impose an upper limit on the number of files in a folder and if necessary split large directories so they do not contain more than a certain maximum number of files (e.g., 10,000). If, for instance, a folder contains one million files then it could instead be organised as a folder (`parent_folder`) with 100 sub-folders (`child00` to `child99`), each containing 10,000 files. This is important because different file systems have different tolerances for handling large numbers of files. For example, the Second Extended Filesystem (ext2) imposes a 'soft' limits of 10,000 files per directory because of the extra overhead when processing such large folders (*The Second Extended Filesystem — The Linux Kernel Documentation*). While modern file systems are capable of handling larger numbers of files, the re-usability of the data will increase when taking into account systems with more modest resources, such as web browsers that may need to list or process all files in a directory.

#### (4) Folder contents.

- a. Group related files unless it is instrumental to keep them separated. For example, group files by specimen, filetype, experimental purpose (treatment, control), etc. It may be crucial to separate different data types into different folders (e.g., one for micrographs and one for particle stacks). Further sub-folders may be necessary for single- and multi-frame micrographs, unaligned and aligned micrographs, etc.
- b. Deposit data from different experimental techniques/modalities as separate archive entries (e.g., single-particle cryoEM data in one, tomography data in another). Some archives allow multiple related but separate entries to be linked or grouped.

### Naming

In this section, we provide some suggestions to improve the naming of files and folders.

#### (5) Meaningful names.

- a. Name files and folders using meaningful identifiers without specifying external references. For instance, while the name `'Figure 5'` probably refers to a figure in a paper describing (some of) the data, users will require access to the article, which may be behind a paywall or in a hard-to-find book. The names of files and folders should exclude any identifiers indicating a particular instrument or your organisation.
- b. Avoid ambiguous attributes such as dates and times particularly in folder names. Mass renaming of files with dates and times can become non-trivial particularly if such attributes vary subtly (e.g., date, minute, seconds) from file to file.

#### (6) Naming symbols.

- a. Restrict names to numerals and lowercase letters and replace all spaces with underscores or hyphens for meaningful word (group) boundaries, to make it substantially easier to work with the data. This facilitates easy transition between typing command line utilities or program names, which invariably work with lowercase (Windows PowerShell cmdlets are case insensitive even though they are documented in CamelCase e.g. `Get-Command`; similarly, macOS path names are case insensitive by default, though this depends on the chosen file-system formatting). Use underscores only for word boundaries and hyphens for keywords or other key attributes such as specimen names identifiable by the presence of a hyphen, e.g., `covid-19`. Consistent use of case also improves readability (Deissenboeck & Pizka, 2006).



- b. Avoid certain characters which could lead to unintended consequences during processing such as ampersands (&), spaces, exclamation marks (!) and question marks (?). In general, stick to the portable character set defined by POSIX and avoid non-ASCII characters (e.g., ü, å or non-Roman scripts) to improve usability. Most keyboards can produce them, and most users will be familiar with them from everyday use. Also, some software will not work with input filenames featuring non-POSIX characters.
- c. Avoid periods in names as this can lead to unpredictable behaviour for instance when attempting to determine formats. For example, while it is generally well known that the file `file.tar.gz` has two standard extensions, it may not be as widely known that `file.ome.tiff`, `file.ome.tf2`, `file.ome.tf8` and `file.ome.btif` are all valid multi-extension bioimaging formats (*OME-TIFF Specification — OME Data Model and File Formats 6.2.2 Documentation*).

### (7) Identity.

- a. Ensure consistency when naming different files and folders related to one another. For example, in [Figure 1](#), labels 6 and 7 show subtle changes in spelling or inclusion/exclusion of characters, which break the naming pattern.
- b. Do not include personal identifiers (e.g., usernames, actual names, etc.) in folder or file names.
- c. Some words to consider for exclusion in the names of files and folders: `'files'`, `'data'`, `'images'` etc. as well as or other words that convey no additional meaningful information.
- d. Think of folder names as applying to all the folders and files they contain as well: there should be no repetition in nested folder names, e.g., `data/control.a/control.a.1/control.a.1.value/data/;`
- e. When providing 3D data as slices or sequentially ordering files, zero-pad the slice/file identifiers correctly (e.g. `prefix-0099.tif` not `prefix-99.tif` for thousands of slices), which guarantees that slices are correctly ordered lexicographically. Failing to do so could result in files being processed in the wrong order and e.g. lead to 3D stacks with misplaced slices, which will affect all analysis steps that follow. For example, consider a volume consisting of 1000 images, each of dimension M by N. Splitting this file should result in file names of the form `file 0001.tif` to `file 1000.tif`. Incorrect names can be fixed using the `rename` shell utility, e.g., `rename file file 00 file?? .tif` will convert all files with 01 to 99 to have 0001 to 0099. `rename` is available on most Linux distributions and may be installed on macOS using Homebrew or from the [source code](#). On Windows systems the [Bulk Rename Utility](#) can be used.

### Miscellaneous

Finally, this section includes some tips on how to handle other aspects of organisation not covered in the previous sections.

### (8) File formats.

- a. Provide images in widely used file formats unless you are demonstrating a novel file format in which case it may be necessary to first get in touch with the archive to plan accordingly. Additional information may be requested to provide users with guidelines on how to use and visualise the new format files including any conversion tools that are available or providing the same data in a widely used file format as well.
- b. Even for file types that are widely used, stick to open formats to ensure that users without access to proprietary software can access the data. Open formats promote the prevalence of tools (open source or proprietary) that can read and write data. We recommend the use of OME-NGFF ([Moore et al., 2021](#)) and OME-TIFF ([Linkert et al., 2010](#)) as open, widely supported imaging file formats.

**(9) Document your data.**

- Include a [README](#) text file which provides an overview of how the data is organised. Depositors may use it to discover the main facets by which the data is organised, the structure of any ad hoc text files as well as the meaning of naming entities used in file/folder names.
- Test the usability of your data by asking a colleague to peruse your data to assess whether the organisation is clear.

**(10) Integrity.** Include checksums, parity codes or hashes for each data file in a separate file, e.g., `md5-sums.txt`, `imageset01.par2` or `sha512-hashes.txt` to facilitate content verification. These will allow users to verify that the data has not been corrupted during the deposition or download process. Each of these different ways to verify file integrity have corresponding tools available for all operating systems, but their operation is beyond the scope of this article (Lianhua & Xingquan, 2017).

Applying the recommendations above, we may revise the path:

`data/A U Thör et al - A very long relevant title that has most of the keywords in your paper/A Folder with an overall description/0923480928 - Treatment Tr1-323 Tissue/0923480928_Treatment_Tr1323_Organelle1-topology1.zip` is '`0923480928 Treatment Tr1-323 Segmentation/0923480928_Treatment_Tr1323_Organelle1-topology1.zip`

to:

`data/brief_description/treatment3_tissue/segmentation/organelle1_topology1.zip`

to achieve a reduction from 328 to 79 characters for the full path. The new organisation is presented in [Figure 2](#).

```

pkorir@x86_64-apple-darwin13:~/PycharmProjects/bandbox
CO2XJWUJGHS P feature/review++ Rg 4.1 kB+ 3.1 kB+
(bandbox) (base) + bandbox git:(feature/review) x tree -I "image*" test_data/0_good
test_data/0_good
├── README.md
├── brief_description
│   ├── control_tissue
│   │   ├── raw
│   │   └── segmentation
│   │       ├── organelle1.zip
│   │       ├── organelle1_topology1.zip
│   │       ├── organelle1_topology2.zip
│   │       ├── organelle2.zip
│   │       ├── organelle3.zip
│   │       ├── organelle4_topology.zip
│   │       └── organelle5_topology.zip
│   ├── treatment1_tissue
│   │   ├── raw
│   │   └── segmentation
│   │       ├── organelle1.zip
│   │       ├── organelle1_topology1.zip
│   │       ├── organelle1_topology2.zip
│   │       ├── organelle2.zip
│   │       ├── organelle3.zip
│   │       ├── organelle4_topology.zip
│   │       └── organelle5_topology.zip
│   ├── treatment2_tissue
│   │   ├── raw
│   │   └── segmentation
│   │       ├── organelle1.zip
│   │       ├── organelle1_topology1.zip
│   │       ├── organelle1_topology2.zip
│   │       ├── organelle2.zip
│   │       ├── organelle3.zip
│   │       ├── organelle4_topology.zip
│   │       └── organelle5_topology.zip
│   └── treatment3_tissue
│       ├── raw
│       └── segmentation
│           ├── organelle1.zip
│           ├── organelle1_topology1.zip
│           ├── organelle1_topology2.zip
│           ├── organelle2.zip
│           ├── organelle3.zip
│           └── organelle4_topology.zip
└── 13 directories, 28 files
(bandbox) (base) + bandbox git:(feature/review) x

```

**Figure 2.** Tree representation of the data from [Figure 1](#) reorganised by applying some of the 10 recommendations described in this paper.

## Conclusion

We hope that these 10 recommendations will only be the beginning of a broader discussion on how to organise bioimaging data in particular and experimental data in general for maximum usefulness, not just to the bioimaging community, but to the wider scientific community. Given the breadth of applications of bioimaging techniques, good organisation would go a long way to helping scientists from other disciplines to benefit from using bioimaging data. There is still considerable scope to develop better ways of not only organising data, but also representing it to enable automated data analysis.

## Data availability

No data are associated with this article.

## Software availability

To make our recommendations practical, we have developed `bandbox` (Korir *et al.*, 2022), an open-source command-line interface (CLI) tool to help users understand how they can improve the organisation of their data in preparation for

```

pkorir@x86_64-apple-darwin13:~/PsycharmProjects/bandbox
(bandbox) (base) → bandbox git:(feature/review) ✖ bandbox analyse test_data/0_bad --prefix test_data/0_bad
misc.                => - unknown file extensions...                               ok
naming               => - accessions in names...                               ok
naming               => - entities with dates in names...           [27 directories] nok
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/0923480928 -
  Treatment1_Tri1-323_Tissue/0923480928_Treatment1-323_Segmentation/0923480928_Treatment1-323_Organelle1-topology1.zip
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Segmentation/9824080293_Control Organelle1 topology1.zip
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/3789392082 -
  Treatment2_Tissue/3789392082_Treatment2_Tissue - Segmentation/3789392082_Treatment2_Organelle1.zip
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9028340982 -
  Treatment1_Tissue/9028340982_Treatment1_Tissue - Segmentation/9028340982_Treatment1_Organelle1.zip
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Segmentation/9824080293_Control Organelle5 topology.zip
  * [+22 other results (include the -a/--all option to view the full list)]
naming               => - excessive periods in names...                               ok
naming               => - long names (>50 chars)...                       [4 directories] nok
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/0923480928 -
  Treatment1_Tri1-323_Tissue/0923480928_Treatment1-323_Segmentation/0923480928_Treatment1-323_Organelle14-topology.zip
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/0923480928 -
  Treatment1_Tri1-323_Tissue/0923480928_Treatment1-323_Segmentation/0923480928_Treatment1-323_Organelle1-topology1.zip
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/0923480928 -
  Treatment1_Tri1-323_Tissue/0923480928_Treatment1-323_Segmentation/0923480928_Treatment1-323_Organelle1-topology2.zip
naming               => - mixed case in names...                               [41 directories] nok
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Raw/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Segmentation/
  * [+36 other results (include the -a/--all option to view the full list)]
naming               => - non-ascii characters in names...           [1 directories] nok
  * A U Thör et al - A very long relevant title that has most of the key words in your paper
naming               => - odd characters [one of '?&! ,'] in names...   [36 directories] nok
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Raw/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Segmentation/
  * [+31 other results (include the -a/--all option to view the full list)]
naming               => - external references in names...                               ok
structure            => - excessives (>2000) files per directory...   ok
structure warning => - obvious directory names...                               [1 directories] nok
  * data/
structure            => - redundant directories...                       [2 directories] nok
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/3789392082 -
  Treatment2_Tissue/3789392082_Treatment2_Tissue - Raw/
structure            => - directories with mixed files...           [1 directories] nok
  * data/A U Thör et al - A very long relevant title that has most of the key words in your paper/A Folder with an overall description/9824080293 -
  Control Tissue/9824080293_Control Tissue - Raw/
    
```

**Figure 3.** Example of a dataset with organisational issues identified by `bandbox` such as use of spaces or non-ASCII characters, redundant directories and other categories with an indication of the number of such entities found in each category. The example dataset is provided with the `bandbox` source code (ASCII – American Standard Code for Information Interchange).



```

pkorir@x86_64-apple-darwin13:~/PycharmProjects/bandbox
(bandbox) (base) + bandbox git:(feature/review) x bandbox analyse test_data/0_good --prefix test_data/0_good
misc. => - unknown file extensions... ok
naming => - accessions in names... ok
naming => - entities with dates in names... ok
naming => - excessive periods in names... ok
naming => - long names (>50 chars)... ok
naming => - mixed case in names... ok
naming => - non-ascii characters in names... ok
naming => - odd characters [one of '&?! ,'] in names... ok
naming => - external references in names... ok
structure => - excessives (>2000) files per directory... ok
structure warning => - obvious directory names... ok
structure => - redundant directories... ok
structure => - directories with mixed files... ok
(bandbox) (base) + bandbox git:(feature/review) x █

```

**Figure 4. Example of a dataset with no issues identified by bandbox.** The example dataset is provided with the [bandbox source code](#).

archival. The program offers two CLI commands: `view` and `analyse`. Running `bandbox view <dir>` command displays a tree of a directory and all its contents; for every non-empty directory with files, `bandbox` provides a summary of the number of files in it, including a list of all the file formats encountered. Running `bandbox analyse <dir>` command provides a listing of possible issues grouped into categories in line with those specified in the Recommendations section. `bandbox` examines the tree associated with the nested hierarchy of files and folders in a dataset and then concurrently runs various heuristics on the tree which are controlled by configurations that the user may modify. The results produced by the `analyse` command are only suggestions for improvement; we understand that there may be practical limitations to implementing some of the suggested improvements as well as good reasons for keeping the data as is. We have designed `bandbox` to be configurable and extensible allowing users to customise analysis parameters (file/folder name length, recognised file formats, accession names, regexes) as well as add new heuristics. An example configuration file is provided in the Github repository. **Figures 3 and 4** show screenshots of the results of running `bandbox` on two different datasets.

Software available from: <https://pypi.org/project/bandbox>

Source code available from: <https://github.com/emdb-empiar/bandbox>

Archived source code at time of publication: <https://doi.org/10.5281/zenodo.7807541> (Korir *et al.*, 2022).

License: [Apache License 2.0](#)

## Acknowledgements

The authors are grateful to Alex J. Noble and Christopher J. Peddie for helpful feedback on the manuscript. We also gratefully acknowledge the many constructive suggestions from the reviewers: S.H.W. Scheres, K.H.L. Ho, S.E. Le Dévédec, W.T. Katz and V. Scarlett. This work aligns with the recommendations of the EuroBioimaging/ELIXIR Joint Strategy ([https://elixir-europe.org/system/files/euro-bioimaging\\_elixir\\_image\\_data\\_strategy.pdf](https://elixir-europe.org/system/files/euro-bioimaging_elixir_image_data_strategy.pdf)), in particular the need for standards and approaches for the organisation of image data storage in established and emerging reference image domains. We acknowledge both ELIXIR and Euro-BioImaging's key roles in highlighting the importance of the effective organisation of biological image data.

## References

- Berman HM, Kleywegt GJ, Nakamura H, *et al.*: **The Protein Data Bank archive as an open data resource.** *J. Comput. Aided Mol. Des.* 2014; **28**(10): 1009–1014.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Brazma A, Hingamp P, Quackenbush J, *et al.*: **Minimum information about a microarray experiment (MIAME)—toward standards for microarray data.** *Nat. Genet.* 2001; **29**(4): 365–371.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Datta S, Lakdawala R, Sarkar S: **Understanding the Inter-Domain Presence of Research Topics in the Computing Discipline.** *IEEE Trans. Emerg. Top. Comput.* 2021; **9**(1): 366–378.  
[Publisher Full Text](#)
- Deissenboeck F, Pizka M: **Concise and consistent naming.** *Softw. Qual. J.* 2006; **14**(3): 261–282.  
[Publisher Full Text](#)
- Ellenberg J, Swedlow JR, Barlow M, *et al.*: **A call for public archives for biological image data.** *Nat. Methods.* 2018; **15**(11): 849–854.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Hartley M, Kleywegt GJ, Patwardhan A, *et al.*: **The BioImage Archive - Building a Home for Life-Sciences Microscopy Data.** *J. Mol. Biol.* 2022; **434**: 167505.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Iudin A, Korir PK, Salavert-Torres J, *et al.*: **EMPIAR: a public archive for raw electron microscopy image data.** *Nat. Methods.* 2016; **13**(5): 387–388.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Iudin A, Korir PK, Somasundharam S, *et al.*: **EMPIAR: the Electron Microscopy Public Image Archive.** *Nucleic Acids Res.* 2023; **51**: D1503–D1511.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Katz WT, Plaza SM: **DVID: Distributed Versioned Image-Oriented Dataservice.** *Front. Neural Circuits.* 2019; **13**.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Korir PK, Iudin A, Somasundharam S, *et al.*: **bandbox (v0.2.1).** *Zenodo.* 2022.  
[Publisher Full Text](#)
- Lianhua C, Xingquan Z: **Hashing Techniques.** *ACM Computing Surveys (CSUR).* 2017.  
[Publisher Full Text](#)
- Li X, Mooney P, Zheng S, *et al.*: **Electron counting and beam-induced motion correction enable near-atomic-resolution single-particle cryo-EM.** *Nat. Methods.* 2013; **10**(6): 584–590.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Linkert M, Rueden CT, Allan C, *et al.*: **Metadata matters: access to image data in the real world.** *J. Cell Biol.* 2010; **189**(5): 777–782.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Mastrorade D: **Tomographic Reconstruction with the IMOD Software Package.** *Microsc. Microanal.* 2006; **12**(S02): 178–179.  
[Publisher Full Text](#)
- Moore J, Allan C, Besson S, *et al.*: **OME-NGFF: a next-generation file format for expanding bioimaging data-access strategies.** *Nat. Methods.* 2021; **18**(12): 1496–1498.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Petek M, Zagorščak M, Blejec A, *et al.*: **pISA-tree - a data management framework for life science research projects using a standardised directory tree.** *Sci. Data.* 2022; **9**(1): 685.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Pietzsch T, Saalfeld S, Preibisch S, *et al.*: **BigDataViewer: visualization and processing for large image data sets.** *Nat. Methods.* 2015; **12**(6): 481–483.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Punjani A, Rubinstein JL, Fleet DJ, *et al.*: **cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination.** *Nat. Methods.* 2017; **14**(3): 290–296.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Rausher MD, McPeck MA, Moore AJ, *et al.*: **Data archiving.** *Evolution.* 2010; **64**(3): 603–604.  
[Publisher Full Text](#)
- Sansone S-A, Rocca-Serra P, Brandizi M, *et al.*: **The first RSBI (ISA-TAB) workshop: "can a simple format work for complex studies?"** *Omic.* 2008; **12**(2): 143–149.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Sarkans U, Chiu W, Collinson L, *et al.*: **REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology.** *Nat. Methods.* 2021; **18**(12): 1418–1422.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Sarkans U, Gostev M, Athar A, *et al.*: **The BioStudies database-one stop shop for all data supporting a life sciences study.** *Nucleic Acids Res.* 2018; **46**(D1): D1266–D1270.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Scheres SHW: **A Bayesian View on Cryo-EM Structure Determination.** *J. Mol. Biol.* 2012; **415**(2): 406–418.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Tang G, Peng L, Baldwin PR, *et al.*: **EMAN2: An extensible image processing suite for electron microscopy.** *J. Struct. Biol.* 2007; **157**(1): 38–46.  
[PubMed Abstract](#) | [Publisher Full Text](#)
- Whitlock MC, McPeck MA, Rausher MD, *et al.*: **Data archiving.** *Am. Nat.* 2010; **175**(2): 145–146.  
[Publisher Full Text](#)
- Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci. Data.* 2016; **3**: 160018.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zhang K: **Gctf: Real-time CTF determination and correction.** *J. Struct. Biol.* 2016; **193**(1): 1–12.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

# Open Peer Review

Current Peer Review Status:    

---

## Version 2

Reviewer Report 09 May 2024

<https://doi.org/10.5256/f1000research.162090.r250664>

© 2024 T. Katz W et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**William T. Katz** 

Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, USA

**Virginia Scarlett** 

<sup>1</sup> Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, Virginia, USA

<sup>2</sup> Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, Virginia, USA

We feel that the article remains at the status of 'Approved with Reservations'. We are grateful to the authors for the revisions that have been implemented, including a clearer explanation of 'organisational resources', improved readability of the figures, and a limit on the output of the bandbox program. However, the article references REMBI but not QUAREP-LiMi, and while it cites OME projects (OME-NGFF and OME-TIFF), it does not mention the impacts or limitations of those important efforts. Also, the article remains vague with respect to its audience. The authors should clarify to which biomedical imaging storage approaches (e.g., flat files, chunked formats, cloud-based APIs, etc.) their recommendations and tooling apply. If the article is intended for users of flat formats (such as TIFF) working on file systems, then the authors should clarify this because the title suggests a very broad applicability to archiving bioimaging data.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Data engineering; biomedical image processing and analysis.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Reviewer Report 04 April 2024

<https://doi.org/10.5256/f1000research.162090.r250665>

© 2024 Ho K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Kenneth H. L. Ho**

Advanced Light Microscopy, The Francis Crick Institute, London, England, UK

The authors have addressed all my previous comments, and I am satisfied with their revision. I have no further comments to make..

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioimage informatics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 18 March 2024

<https://doi.org/10.5256/f1000research.162090.r250663>

© 2024 Le Dévédec S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Sylvia Emmanuelle Le Dévédec** 

Division of Drug Discovery and Safety, Leiden Academic Centre of Drug Research, Universiteit Leiden, Leiden, South Holland, The Netherlands

The revised version has effectively addressed the majority of my comments. Clarity regarding the intended audience and the importance of data archiving has been improved, and ambiguities have been identified and addressed.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Biology; image-based phenotypic profiling; microscopist; data generator; core facility management; FAIR metadata

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Reviewer Report 13 March 2024

<https://doi.org/10.5256/f1000research.162090.r250662>

© 2024 Scheres S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Sjors Scheres** 

Medical Research Council Laboratory of Molecular Biology, Cambridge, England, UK

Most of my comments have been addressed in the revised version.

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Structural biologist; software developer

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---

### Version 1

Reviewer Report 21 November 2023

<https://doi.org/10.5256/f1000research.142422.r217552>

© 2023 T. Katz W et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**William T. Katz** 

Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, USA

**Virginia Scarlett** 

<sup>1</sup> Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, Virginia, USA

<sup>2</sup> Howard Hughes Medical Institute's Janelia Research Campus, Ashburn, Virginia, USA

In this opinion article, the authors tackle an important but often-overlooked aspect of biomedical data archives: how best to organize data folders and files to maximize ease of use. Ten recommendations are provided as well as a lightweight command-line tool for inspecting datasets. Since there continues to be an acceleration in both the number and size of these datasets accessible through various repositories, both the recommendations and tool from experienced archivists are useful and should be published, though we feel some revision of the document is warranted.

The introduction describes the broader context of bioimaging data management before focusing on the contributions of the article. There could be clearer differentiation of efforts to standardize bioimaging metadata (REMBI, QUAREP-LiMi), file formats and associated libraries (OME-TIFF, OME-NGFF, Zarr, n5), and local or cloud-based services that provide Data APIs (DVID, BossDB) with



some level of abstraction in how data is actually stored. The recommendations and tool mainly apply to file-based solutions though some of the recommendations, such as naming, would be applicable to other forms of big data repositories. We suggest that the authors clarify the scope of their contributions.

It should be noted that some of the efforts to standardize data and its distribution also have recommendations for organization of data. For example, OME-NGFF requires segmentation to be in a directory called "labels/".

In the third paragraph of Motivation, the terms "ways and means" and "organisational resources" are unclear though some of your examples (folder hierarchy, file formats, identifiers) show how data can be organized. We suggest you start with some examples and then introduce "organizational resources" as a term.

If standardization is not an aim, can bandbox be configured to remove warnings not agreed upon by a user? In Figure 2, the printing of the word "warning" for datasets with no red flags seems odd. We would suggest using "check" as in "name check" or "structure check" if no warnings exist.

Given recommendation (8)b and the article's bioimaging focus, the bandbox tool should work by default with well-known, large-scale formats like Zarr and N5. In testing, it appears that bandbox doesn't recognize file extensions used by such formats like .json and .zarr. The configurability of bandbox is a nice feature and should be mentioned in the article. This would allow other tool builders to contribute configurations for validating common formats and it seems like the regex capability could allow folder hierarchy requirements.

The command-line bandbox tool should limit warning output to some maximum number of lines by default. This is particularly true for massive, chunked datasets consisting of many files and folders. We would suggest adding a "verbose" flag to allow full results to be output perhaps to a file.

Some minor points:

The description of the bandbox tool could be moved out of the Motivation section and after listing the recommendations.

Figure 1 has too small font sizes and would not be readable for printed copies as well as expending quite a bit of black ink.

The phenomenon described in the first sub-bullet under 'Verbosity' is an interesting point that seems to deserve its own name. Maybe something like, 'redundant nesting' or 'over-nesting'. A name would also make it easier to connect to solution 3A, which is conceptually related. Also, the second half of this bullet point is in monospace font, but it should be Times New Roman or whatever.

Some recommendations are more universally advisable than others. For those points, we'd recommend dropping "Consider" for stronger language.

In (4)b, "Most archives allow multiple separate entries to be linked or grouped," it's not clear what

qualifies as an archive since data could be made available through cloud providers' object stores and other facilities.

In (5)b, could you clarify in what ways dates and times are "ambiguous attributes"?

In the sentence, "bandbox examines the tree associate with the nested hierarchy..." the word bandbox should be in monospace font.

What is the rationale for limiting folder depth to 3 or 4 levels?

In (7)b, "Do not include personal identifiers in folder names." Personal identifiers should be clarified.

For (7)e, zero-padding should be considered for any sequentially ordered set of files. A good case is 2D slices of a 3D volume as described.

For (8)b, consider citing OME-NGFF and OME-TIFF as recommended community formats.

For (9)a, the recommendation for an overview could explicitly suggest listing the facets used to organize the data.

In Figure 4, is the single "brief\_description" folder at that level recommended instead of adding the descriptive information to a README file? Perhaps a real description should be used in the example to make it clear why recommendation (3)b doesn't apply.

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Partly

**Are arguments sufficiently supported by evidence from the published literature?**

Partly

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Data engineering; biomedical image processing and analysis.

**We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.**

Author Response 30 Jan 2024

**Gerard J Kleywegt**

**Response to Reviewer #4 (in italics)**

Specific comments:

In this opinion article, the authors tackle an important but often-overlooked aspect of biomedical data archives: how best to organize data folders and files to maximize ease of use. Ten recommendations are provided as well as a lightweight command-line tool for inspecting datasets. Since there continues to be an acceleration in both the number and size of these datasets accessible through various repositories, both the recommendations and tool from experienced archivists are useful and should be published, though we feel some revision of the document is warranted.

The introduction describes the broader context of bioimaging data management before focusing on the contributions of the article. There could be clearer differentiation of efforts to standardize bioimaging metadata (REMBI, QUAREP-LiMi), file formats and associated libraries (OME-TIFF, OME-NGFF, Zarr, n5), and local or cloud-based services that provide Data APIs (DVID, BossDB) with some level of abstraction in how data is actually stored. The recommendations and tool mainly apply to file-based solutions though some of the recommendations, such as naming, would be applicable to other forms of big data repositories. We suggest that the authors clarify the scope of their contributions.

It should be noted that some of the efforts to standardize data and its distribution also have recommendations for organization of data. For example, OME-NGFF requires segmentation to be in a directory called "labels/".

In the third paragraph of Motivation, the terms "ways and means" and "organisational resources" are unclear though some of your examples (folder hierarchy, file formats, identifiers) show how data can be organized. We suggest you start with some examples and then introduce "organizational resources" as a term.

*We accept this correction and have updated the text to better reflect this point.*

If standardization is not an aim, can bandbox be configured to remove warnings not agreed upon by a user? In Figure 2, the printing of the word "warning" for datasets with no red flags seems odd. We would suggest using "check" as in "name check" or "structure check" if no warnings exist.

*We have released an updated version (bandbox v0.2.2) where these have been amended.*

Given recommendation (8)b and the article's bioimaging focus, the bandbox tool should work by default with well-known, large-scale formats like Zarr and N5. In testing, it appears that bandbox doesn't recognize file extensions used by such formats like .json and .zarr. The configurability of bandbox is a nice feature and should be mentioned in the article. This would allow other tool builders to contribute configurations for validating common formats and it seems like the regex capability could allow folder hierarchy requirements.

*We have clarified in the text that bandbox is configurable.*

The command-line bandbox tool should limit warning output to some maximum number of lines by default. This is particularly true for massive, chunked datasets consisting of many files and folders. We would suggest adding a “verbose” flag to allow full results to be output perhaps to a file.

*This has been updated in bandbox v0.2.2. Instead of printing all results by default, we have substituted the -S/--summarise flag with a -a/--all flag so that by default users don't get overwhelmed. The instruction to use the new flag is now highlighted in yellow text beneath each section with more than a certain number of results.*

Some minor points:

The description of the bandbox tool could be moved out of the Motivation section and after listing the recommendations.

*We have now included a detailed description of bandbox in the Software Availability section.*

Figure 1 has too small font sizes and would not be readable for printed copies as well as expending quite a bit of black ink.

*We accept the suggestion and have changed all images to have a light background.*

The phenomenon described in the first sub-bullet under 'Verbosity' is an interesting point that seems to deserve its own name. Maybe something like, 'redundant nesting' or 'over-nesting'. A name would also make it easier to connect to solution 3A, which is conceptually related. Also, the second half of this bullet point is in monospace font, but it should be Times New Roman or whatever.

*We have given the section the name 'Verbosity/Redundancy'.*

*The use of monospace font here is intentional to distinguish between literal text and computer text (file/folder names, commands, tools).*

Some recommendations are more universally advisable than others. For those points, we'd recommend dropping “Consider” for stronger language.

In (4)b, “Most archives allow multiple separate entries to be linked or grouped,” it's not clear what qualifies as an archive since data could be made available through cloud providers' object stores and other facilities.

*We have provided a definition of 'archive' in the opening paragraph of the article.*

In (5)b, could you clarify in what ways dates and times are “ambiguous attributes”?

*We have provided an explanation on this in the text.*

*Dates and times are ambiguous to the extent that they do not provide meaningful attributes associated with the experiment. While it can be assumed that dates on file names refer to the date of collection, this is not instrumental to the actual data i.e. knowing the date of collection adds no scientific value. Furthermore, having every single image file with the same date consumes precious 'naming space' of files, which can either be provided once in the name of the parent folder or as part of the metadata, where it would be expected to convey useful information to users.*

In the sentence, "bandbox examines the tree associate with the nested hierarchy..." the word bandbox should be in monospace font.

*This has been corrected in the text.*

What is the rationale for limiting folder depth to 3 or 4 levels?

*We have argued this point based on the ISA framework.*

In (7)b, "Do not include personal identifiers in folder names." Personal identifiers should be clarified.

*We accept this point and have provided some examples of what is meant by 'personal identifiers'.*

For (7)e, zero-padding should be considered for any sequentially ordered set of files. A good case is 2D slices of a 3D volume as described.

*We have included 'sequential ordering' as another example of this phenomenon.*

For (8)b, consider citing OME-NGFF and OME-TIFF as recommended community formats.

*We accept the suggestion and have amended the text as requested.*

For (9)a, the recommendation for an overview could explicitly suggest listing the facets used to organize the data.

*We have included a sentence outlining what may be included in the README file.*

In Figure 4, is the single "brief\_description" folder at that level recommended instead of adding the descriptive information to a README file? Perhaps a real description should be used in the example to make it clear why recommendation (3)b doesn't apply.

*This term is purely illustrative as are the names of the files and folders.*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 21 November 2023

<https://doi.org/10.5256/f1000research.142422.r217556>

© 2023 Le Dévédec S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

? **Sylvia Emmanuelle Le Dévédec** 

Division of Drug Discovery and Safety, Leiden Academic Centre of Drug Research, Universiteit Leiden, Leiden, South Holland, The Netherlands

The guideline presented by Korir and colleagues, who are recognized experts in data structure, data organization, and FAIRification, marks an important step toward fostering a comprehensive discussion on the management of bioimaging data. The authors, primarily developers and bioinformaticians dealing with intricate datasets, have assembled a set of recommendations that, while valuable, may be perceived as overly abstract, potentially posing challenges for experimentalists who serve as the primary data producers.

One critical aspect that emerges is the need for greater clarity regarding the intended audience for this guideline. Currently, it appears somewhat ambiguous, leading to potential misalignment with the individuals it should be primarily targeting. It is recommended that the authors explicitly define the community they aim to address at the outset of the manuscript. If, indeed, the target audience is the data producers, particularly experimentalists, then a comprehensive revision of the recommendations may be necessary. Consideration should be given to conveying the guidelines in a more accessible language, ensuring that the practical implications for experimentalists are clearly delineated. Additionally, the authors might explore the possibility of tailoring specific sets of guidelines for distinct roles, such as data managers and data producers, to enhance relevance and utility.

Below are listed some specific points of attentions:

Abstract:

- The abstract lacks clarity on the intended audience of these recommendations. It is essential to specify whether the guidelines primarily target core facility managers, data managers/stewards, bioinformaticians, or experimentalists.
- If the guidelines are intended for experimentalists, the current manuscript may not align with the needs of this non-expert audience. The language and content may need to be adapted to cater to individuals with limited knowledge in data management.
- The phrase "make future data depositions more useful" needs clarification. Who benefits from this increased usefulness, and in what way? Is the goal to enhance practicality, efficiency, or accessibility? A more specific explanation would enhance the abstract's clarity.
- The term "bioimaging community" is used in the abstract, but its specific meaning in the context of this manuscript is unclear. Defining this community will provide readers with a

better understanding of the scope and relevance of the guidelines.

- The abstract mentions that Bandbox is designed "to facilitate the process of analyzing data organization." It would be beneficial to elaborate on how the analyzing functionality of Bandbox directly benefits the bioimaging community. Specific examples or scenarios demonstrating its advantages would enhance the abstract's informativeness.

#### Introduction

- What does data 'archiving' means exactly in this specific manuscript?
- Objective of the guideline: Harmonising how to organise datasets for maximum usefulness with archival in mind?
- Where this organization should occur in the data life cycle: before/during or after generation? Where this organisation should occur? In which physical storage space?
- Organisation = order of the data. Organisation or order of the data should be implicitly connected to the related metadata and even contained somewhere in the metadata.
- 'Good organisation (order) of data improves its usefulness and is the responsibility of the data depositors.' Do you mean here the data generator or specifically the data depositor? Based on the description it seems like the data depositor is implicitly the data generator.
- 'Users can immediately distinguish the various experimental categories': should you not refer to (p)ISA to clarify what is meant by 'experimental categories' (<https://doi.org/10.1038/s41597-022-01805-5>)?
- 'Facet refers to the various attributes germane to the experiment which may be included in the folder and file names'. Should 'facet' not be called 'key'? If not then explain the differences between both terms.

#### Recommendations:

- The potential users for these recommendations lack clear definition, and depending on the proposed users, the guide should be tailored for optimal understanding. Data depositors and generators often have different levels of familiarity compared to program developers or data stewards, employing distinct languages. Addressing these differences is crucial for ensuring accessibility and effectiveness.
- Open-source command-line interfaces can be intimidating, particularly for experimentalists who serve as the primary data generators and often act as data depositors. As a cell biologist and experimentalist, I find the proposed CLI tool, while impressive and useful, potentially challenging to navigate comfortably. Enhancements in user-friendliness or alternative interfaces might significantly benefit experimentalists who are integral to both data generation and deposition.
- Given the recommendation for data producers to pre-define structures before data collection, it becomes apparent that the target audience of this guide is experimentalists with limited knowledge of data management and programming. Including guidelines or tips on naming conventions would be particularly valuable for such users, enhancing the practicality and applicability of the recommendations.

- The suggestion regarding folder contents description appears somewhat vague and may not be universally suitable for various experiment types. A more nuanced approach that considers the diversity of experiments would enhance the guide's usability.
- The concept of "meaningful names" for folders raises questions about subjectivity and human sensitivity, which may not align with the precision required for effective data management structures. Establishing a clear naming convention, is objectively applicable across various contexts, would contribute to the robustness and reliability of the guide.

## References

1. Petek M, Zagorščak M, Blejec A, Ramšak Ž, et al.: pISA-tree - a data management framework for life science research projects using a standardised directory tree. *Scientific Data*. 2022; **9** (1).  
[Publisher Full Text](#)

### Is the topic of the opinion article discussed accurately in the context of the current literature?

Partly

### Are all factual statements correct and adequately supported by citations?

Yes

### Are arguments sufficiently supported by evidence from the published literature?

Partly

### Are the conclusions drawn balanced and justified on the basis of the presented arguments?

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Biology; image-based phenotypic profiling; microscopist; data generator; core facility management; FAIR metadata

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 30 Jan 2024

**Gerard J Kleywegt**

#### **Response to Reviewer #3 (in italics)**

Specific comments:

The guideline presented by Korir and colleagues, who are recognized experts in data



structure, data organization, and FAIRification, marks an important step toward fostering a comprehensive discussion on the management of bioimaging data. The authors, primarily developers and bioinformaticians dealing with intricate datasets, have assembled a set of recommendations that, while valuable, may be perceived as overly abstract, potentially posing challenges for experimentalists who serve as the primary data producers.

One critical aspect that emerges is the need for greater clarity regarding the intended audience for this guideline. Currently, it appears somewhat ambiguous, leading to potential misalignment with the individuals it should be primarily targeting. It is recommended that the authors explicitly define the community they aim to address at the outset of the manuscript. If, indeed, the target audience is the data producers, particularly experimentalists, then a comprehensive revision of the recommendations may be necessary. Consideration should be given to conveying the guidelines in a more accessible language, ensuring that the practical implications for experimentalists are clearly delineated. Additionally, the authors might explore the possibility of tailoring specific sets of guidelines for distinct roles, such as data managers and data producers, to enhance relevance and utility.

Below are listed some specific points of attentions:

Abstract:

- The abstract lacks clarity on the intended audience of these recommendations. It is essential to specify whether the guidelines primarily target core facility managers, data managers/stewards, bioinformaticians, or experimentalists.

*We welcome the suggestion to clarify the intended audience and have updated the abstract to clarify this.*

- If the guidelines are intended for experimentalists, the current manuscript may not align with the needs of this non-expert audience. The language and content may need to be adapted to cater to individuals with limited knowledge in data management.

*We aim to address a wide and varied audience, so the language and terminology needs to strike a balance for the content to be accessible and digestible by different groups. We hope we have managed a reasonable balance, especially following the many constructive suggestions of all the reviewers. If this reviewer has specific comments on sections in the revised manuscript that could be improved further in this respect we would be happy to attempt to do so.*

- The phrase "make future data depositions more useful" needs clarification. Who benefits from this increased usefulness, and in what way? Is the goal to enhance practicality, efficiency, or accessibility? A more specific explanation would enhance the abstract's clarity.

*We accept the correction and have spelled out in more precise terms what 'more useful' means and to whom this applies.*

- The term "bioimaging community" is used in the abstract, but its specific meaning in the context of this manuscript is unclear. Defining this community will provide readers with a better understanding of the scope and relevance of the guidelines.

*We accept the correction and have amended the text to reflect this.*

- The abstract mentions that Bandbox is designed "to facilitate the process of analyzing data organization." It would be beneficial to elaborate on how the analyzing functionality of Bandbox directly benefits the bioimaging community. Specific examples or scenarios demonstrating its advantages would enhance the abstract's informativeness.

*We accept the correction and have included, in the text, some examples of what bandbox is capable of doing.*

#### Introduction

- What does data 'archiving' mean exactly in this specific manuscript?

*We have included a definition of 'archiving' in the opening paragraph of the article.*

- Objective of the guideline: Harmonising how to organise datasets for maximum usefulness with archival in mind?

*Yes.*

- Where this organization should occur in the data life cycle: before/during or after generation? Where this organisation should occur? In which physical storage space?

*The earlier the better. Recommendation #1 (Design before data collection) highlights the impact of data planning before collection commences. The remaining recommendations outline various suggestions on how to improve the usability of the data. The organisation typically would happen on the storage device but can be done either through consoles or the appropriate graphical user interfaces.*

- Organisation = order of the data. Organisation or order of the data should be implicitly connected to the related metadata and even contained somewhere in the metadata.

*We have edited the text for clarity.*

- 'Good organisation (order) of data improves its usefulness and is the responsibility of the data depositors.' Do you mean here the data generator or specifically the data depositor? Based on the description it seems like the data depositor is implicitly the data generator.

*Data depositors' here refers to the individual(s) responsible for making the submission to the archive (previously defined) and this may or may not be the generator of the data. In*

*many cases, the depositor is familiar with the data because they performed the analyses implying familiarity with handling the data.*

- ‘Users can immediately distinguish the various experimental categories’: should you not refer to (p)ISA to clarify what is meant by ‘experimental categories’ (<https://doi.org/10.1038/s41597-022-01805-5>)?

*We are grateful to the reviewer for pointing out this reference which is now referred to in the text.*

- ‘Facet refers to the various attributes germane to the experiment which may be included in the folder and file names’. Should ‘facet’ not be called ‘key’? If not then explain the differences between both terms.

*We used the term ‘facet’ in the same sense as in multifaceted, implying that a dataset may be viewed from various perspectives to discern distinct properties much in the same way as a gem. The reviewer’s proposal of ‘key’ does not fit this sense.*

#### Recommendations:

- The potential users for these recommendations lack clear definition, and depending on the proposed users, the guide should be tailored for optimal understanding. Data depositors and generators often have different levels of familiarity compared to program developers or data stewards, employing distinct languages. Addressing these differences is crucial for ensuring accessibility and effectiveness.

*We have addressed the specificity of the audience in the amendments to the abstract (above).*

- Open-source command-line interfaces can be intimidating, particularly for experimentalists who serve as the primary data generators and often act as data depositors. As a cell biologist and experimentalist, I find the proposed CLI tool, while impressive and useful, potentially challenging to navigate comfortably. Enhancements in user-friendliness or alternative interfaces might significantly benefit experimentalists who are integral to both data generation and deposition.

*We accept this comment and are only constrained by our capacity to extend the CLI tool to achieve the desired usability.*

- Given the recommendation for data producers to pre-define structures before data collection, it becomes apparent that the target audience of this guide is experimentalists with limited knowledge of data management and programming. Including guidelines or tips on naming conventions would be particularly valuable for such users, enhancing the practicality and applicability of the recommendations.

*Recommendations 5, 6 and 7 go into considerable detail about what names to choose, which symbols to use in names and matters relating to identity. We are willing to revise any of the provided recommendations which remain unclear.*

- The suggestion regarding folder contents description appears somewhat vague and may not be universally suitable for various experiment types. A more nuanced approach that considers the diversity of experiments would enhance the guide's usability.

*We appreciate that the authorship of this article does not represent the universe of experimental methods in imaging. We do point out various facets that may be relevant but leave it up to depositors (generators) who are in the best position to judge which to use when structuring/naming folders. We also point out in the abstract that we offer these recommendations to start discussions in various data-rich communities.*

- The concept of "meaningful names" for folders raises questions about subjectivity and human sensitivity, which may not align with the precision required for effective data management structures. Establishing a clear naming convention, is objectively applicable across various contexts, would contribute to the robustness and reliability of the guide.

*As stated above, we do not think it necessary to specify exactly how data should be organised given the vast variety of experiments that can be carried out. We do state in the article (Motivation, paragraph 8) that "...our guide is intended to lead towards best practices rather than serve as a framework. ...this guide does not aim to achieve standardisation. We believe it is more practical to have a set of best practices and leave it up to the data authors to decide how best to apply them."*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 08 November 2023

<https://doi.org/10.5256/f1000research.142422.r217555>

© 2023 Ho K. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Kenneth H. L. Ho**

Advanced Light Microscopy, The Francis Crick Institute, London, England, UK

The article is timely as we are facing a deluge of bioimaging data with higher resolutions and automation. It is therefore an area that needs more discussions, sharing of good practices.

I mostly agree with all the recommendations given, although I feel that the authors may need to make a good argument for some recommendations. Some choices seem arbitrary and I would like to see the rationale behind them.

After reading the paper, I am a bit confused by the article's intended target audience. Is the article recommendation aimed at most of the biologists who archive their bioimaging data mainly for the purpose of peer review and references? Or is the article and recommendation aiming for those database curators and producers of bioimaging databases, e.g. IDR (<https://idr.openmicroscopy.org/>), SSBD (<https://ssbd.riken.jp/database/>), GDC (<https://portal.gdc.cancer.gov/>), etc?

On page 4 under the heading 'Motivation', "We believe that recommendations outlined here maybe of value to two principal groups of users: 1) data depositors, who need to design and prepare their data to improve its usability to the community".

Does it include most biologists? I believe that most biologists archive their data to provide a record of their studies. Are the data depositors in the article and its intended audience refer to bioimaging database curators/producers instead of bench biologists?

I believe that the ten recommendations would be equally applied to most biologists even though their aim is to provide a record of their studies, the recommendations would help those database curators to organise their bioimage data in more meaningful ways.

I would like to see that part to be make clearer of its intended audience.

With regards to the recommendations, on page 8, 'Naming' (5) Meaningful names (b) "Consider avoiding ambiguous attributes such as dates and times.

The argument that they have "subtle variations" is not obvious to me. Is it because of variations of date formats used in different countries? Would it be solved if ISO 8601 ([https://en.wikipedia.org/wiki/ISO\\_8601](https://en.wikipedia.org/wiki/ISO_8601)) date format is used? Would that be a better recommendation? If not, would the authors care to expand their argument for that as dates are used frequently in filenames?

On page 9, (6) Naming symbols, (a) consider confining to lowercase letters.

It seems to be rather arbitrary to confine names to lowercase, why would it not work for all uppercase letters instead?

Similarly, in (b) avoid non-ASCII characters. Shouldn't we be more inclusive of other languages that are non-ascii, e.g., European characters, or double byte Japanese, Korean and Chinese characters?

From a computer coding point of view, I intuitively understand the rationale for choosing ASCII but the article doesn't seem to provide a valid argument for it. May I suggest the authors to use international standard for POSIX Portable Operating System Interface (IEEE 1003 ISO/IEC 9945) (ref: <https://en.wikipedia.org/wiki/POSIX>; <https://www.ibm.com/docs/en/zos/2.2.0?topic=locales-posix-portable-file-name-character-set>) instead. Choosing to use an international standard makes more sense instead of creating another separate standard specifically for bioimaging data. If the authors would like to keep their recommendations, I would like to see more justification for doing so.

On (d) upper limit on the length of file and folder names. The authors proposed a working upper limit of 50 characters. Again, it seems to be arbitrary, why not 80 characters, i.e. one line length on the old CRT terminal? The browser limit is a good reason, but I would like to see a more robust

argument that 50 characters length is a good compromise.

The authors used an example of file path limit of 320 characters in the same paragraph, I believe it may cause confusion for the reader with filename length, which for most computer systems, is only 255 characters. (ref: <https://en.wikipedia.org/wiki/Filename> ). Since the authors also provide recommendation (3) on Folder depth and given example of path length problems on page 7 “Very long names of files/folders”, maybe the authors can discuss and recommend that together, under one section “filename length, path length and folder depth”. It may be easier for the reader to appreciate the choice that the authors make.

On recommendation (8) Friendly file formats. Maybe “Widely used file formats” is more applicable? I would prefer “Openly accessible file formats”, i.e., formats that there are readable by open-source tools. I guess widely used file formats would fit that description too and reflect more closely to what the authors want to convey. Proprietary software tools for accessing proprietary file formats may cause problems in the long run as companies often change hands, e.g., Olympus is now Evident, LaVision is now under Bruncker. It is difficult to ensure that companies will keep supporting certain formats in their software tools in the future while funding bodies (in the UK) require archiving data for 10 to 20 years.

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Yes

**Are arguments sufficiently supported by evidence from the published literature?**

Partly

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Bioimage informatics

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 30 Jan 2024

**Gerard J Kleywegt**

**Response to Reviewer #2 (in italics)**

Specific comments:

The article is timely as we are facing a deluge of bioimaging data with higher resolutions and automation. It is therefore an area that needs more discussions, sharing of good practices.

I mostly agree with all the recommendations given, although I feel that the authors may need to make a good argument for some recommendations. Some choices seem arbitrary and I would like to see the rationale behind them.

After reading the paper, I am a bit confused by the article's intended target audience. Is the article recommendation aimed at most of the biologists who archive their bioimaging data mainly for the purpose of peer review and references? Or is the article and recommendation aiming for those database curators and producers of bioimaging databases, e.g. IDR (<https://idr.openmicroscopy.org/>), SSBD (<https://ssbd.riken.jp/database/>), GDC (<https://portal.gdc.cancer.gov/>), etc?

On page 4 under the heading 'Motivation', "We believe that recommendations outlined here maybe of value to two principal groups of users: 1) data depositors, who need to design and prepare their data to improve its usability to the community".

Does it include most biologists? I believe that most biologists archive their data to provide a record of their studies. Are the data depositors in the article and its intended audience refer to bioimaging database curators/producers instead of bench biologists?

I believe that the ten recommendations would be equally applied to most biologists even though their aim is to provide a record of their studies, the recommendations would help those database curators to organise their bioimage data in more meaningful ways.

I would like to see that part to be make clearer of its intended audience.

*We accept the correction and have expanded the introductory paragraphs to outline specific audiences as well as clarified the type of user that 'user' refers to.*

With regards to the recommendations, on page 8, 'Naming' (5) Meaningful names (b) "Consider avoiding ambiguous attributes such as dates and times.

The argument that they have "subtle variations" is not obvious to me. Is it because of variations of date formats used in different countries? Would it be solved if ISO 8601 ([https://en.wikipedia.org/wiki/ISO\\_8601](https://en.wikipedia.org/wiki/ISO_8601)) date format is used? Would that be a better recommendation? If not, would the authors care to expand their argument for that as dates are used frequently in filenames?

*We accept the correction and have edited the text to better reflect the intended meaning.*

On page 9, (6) Naming symbols, (a) consider confining to lowercase letters.

It seems to be rather arbitrary to confine names to lowercase, why would it not work for all uppercase letters instead?

*We accept the correction and include arguments why we think it is preferable for file and folder*

*names to be defined using lowercase letters.*

Similarly, in (b) avoid non-ASCII characters. Shouldn't we be more inclusive of other languages that are non-ascii, e.g., European characters, or double byte Japanese, Korean and Chinese characters?

From a computer coding point of view, I intuitively understand the rationale for choosing ASCII but the article doesn't seem to provide a valid argument for it. May I suggest the authors to use international standard for POSIX Portable Operating System Interface (IEEE 1003 ISO/IEC 9945) (ref: <https://en.wikipedia.org/wiki/POSIX>; <https://www.ibm.com/docs/en/zos/2.2.0?topic=locales-posix-portable-file-name-character-set>) instead. Choosing to use an international standard makes more sense instead of creating another separate standard specifically for bioimaging data. If the authors would like to keep their recommendations, I would like to see more justification for doing so.

*We accept the correction and now refer to POSIX as the standard to adhere to as well as provide reasons to do so.*

On (d) upper limit on the length of file and folder names. The authors proposed a working upper limit of 50 characters. Again, it seems to be arbitrary, why not 80 characters, i.e. one line length on the old CRT terminal? The browser limit is a good reason, but I would like to see a more robust argument that 50 characters length is a good compromise.

*The reviewer's comment does raise a valid point. However, it is important to bear in mind that file and folder names add to one another and a length of 80 means that at a depth of three folders will admit paths of up to 240 characters. It is hard to precisely determine what would be reasonable: 20-30 characters may be too short for a lot of cases. One option would be to examine file lengths in current archives to determine the distribution of file name lengths but if the objective is to follow good rather than current practice this may not be sound.*

*The authors propose the above limits to start a conversation with the community on what would be a sensible value or range.*

The authors used an example of file path limit of 320 characters in the same paragraph, I believe it may cause confusion for the reader with filename length, which for most computer systems, is only 255 characters. (ref: <https://en.wikipedia.org/wiki/Filename> ). Since the authors also provide recommendation (3) on Folder depth and given example of path length problems on page 7 "Very long names of files/folders", maybe the authors can discuss and recommend that together, under one section "filename length, path length and folder depth". It may be easier for the reader to appreciate the choice that the authors make.

*We accept the correction and have restructured the article as suggested.*

On recommendation (8) Friendly file formats. Maybe "Widely used file formats" is more applicable? I would prefer "Openly accessible file formats", i.e., formats that there are readable by open-source tools. I guess widely used file formats would fit that description



too and reflect more closely to what the authors want to convey. Proprietary software tools for accessing proprietary file formats may cause problems in the long run as companies often change hands, e.g., Olympus is now Evident, LaVision is now under Brunner. It is difficult to ensure that companies will keep supporting certain formats in their software tools in the future while funding bodies (in the UK) require archiving data for 10 to 20 years.

*We have revised the section title to simply 'File formats'. We appreciate that there are file formats that are unavoidable but proprietary (e.g., from microscopes) but our emphasis is on the openness of the formats because this enables the prevalence of tools which can reliably read the data. We have updated 8(b) to reflect this point.*

**Competing Interests:** No competing interests were disclosed.

Reviewer Report 01 November 2023

<https://doi.org/10.5256/f1000research.142422.r217553>

© 2023 Scheres S. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



**Sjors Scheres** 

Medical Research Council Laboratory of Molecular Biology, Cambridge, England, UK

This paper describes recommendations for organizing imaging data from the life sciences for archival purposes. Coming from the EBI, which is responsible for a large proportion of image archiving in the field, this advice is important and worth of dissemination to the wider scientific community. I am therefore, in principle, enthusiastic about its publication in F1000Research. However, I do think that the manuscript and the explicit recommendations can be improved, as the phrasing is often vague and some of the recommendations are ignored by the authors themselves. I would therefore recommend a careful re-think and re-write, especially of the 10 recommendations, for a revised version.

Specific comments:

**Abstract:**

p1: The first sentence does not make sense: is 'organised data' an elusive goal?

Motivation:

p3: Would you not consider non-scientists looking at these images?

p3: "in the use of 'ways and means' of effecting the organisation"  
-> I have no clue what this means.

p4: "To achieve this ... and so on"

-> These vague statements need rephrasing (e.g. 'we define [...] to the \*various\* attributes'). Also, what is 'generally available equipment'?

p5: it is not entirely clear to me from reading the paper what the bandbox program does. The paper states that it is based on the 10 recommendations that follow, but as explained below the re-organisation in Figure 4 still violates several recommendations... Perhaps some pseudo-code may be useful? Also, wouldn't it make more sense to first describe the recommendations and then introduce this program?

### **Recommendations:**

Except for recommendation (7), all recommendations start with the word 'consider'. Given these are recommendations, that is superfluous. It may be clearer to use an imperative to directly state the recommendation (like done in 7).

p6: How is a "raw" TIFF file defined?

(3a) "the fewer the better" means a depth of 1 is best. This is probably not what the authors intended.

(3b) Having a subfolder called 'tiff' is often a good idea, e.g. when there is also a file with metadata describing those tiff images (which is typically the case). In fact, the recommended Figure 4 has a 'raw' folder, which has exactly the same meaning, thus contradicting this recommendation.

(5a) What are "any references that are tied to the instrument" and why should they be excluded? If these are references to the microscope used, they may be relevant to the user?

(5b) Why would dates in filenames be ambiguous and should they be avoided? Many data acquisition softwares write files with date and times in their names. Renaming these would, as the authors themselves point out, indeed be complicated and possibly lead to errors.

(7a) I have no clue what this means: "similar folders at different depths have the same names"

(7b) What are "personal identifiers"?

(7c) The name 'data' is actually used in the line below and in the recommended Figure 4. Also, I don't see why 'images' won't be an excellent name for a folder that contains images?

(7e) This recommendation may not be limited to slices of 3D data, which seems an arbitrarily narrow example for such broad recommendations. I personally thought of zer-padding images when I first read this (apparently not careful enough!). Perhaps using a term like "leading zeros" may be less ambiguous? Albeit perhaps useful to some of the readers, this is the only recommendation that has an explicit explanation of how to do this on two specific computer systems. Wouldn't this be something that could only be implemented in the bandbox program, so it could be used on any computer?

(8) What are "friendly file formats?". Also, the term "widely used file formats" is not well defined.

(9a) The example in Figure 4 does not have a README file...

(9b) "This can be achieved ... data presents" -> This sounds superfluous and condescending.

p11: The proposed path "data/brief\_description/treatment3..." violates at least recommendations 3 (unused subfolder 'brief\_description') and 7 (use of word 'data')

**Is the topic of the opinion article discussed accurately in the context of the current literature?**

Yes

**Are all factual statements correct and adequately supported by citations?**

Partly

**Are arguments sufficiently supported by evidence from the published literature?**

Yes

**Are the conclusions drawn balanced and justified on the basis of the presented arguments?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Structural biologist; software developer

**I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.**

Author Response 30 Jan 2024

**Gerard J Kleywegt**

**Response to Reviewer #1 (in italics)**

Specific comments:

**Abstract:**

p1: The first sentence does not make sense: is 'organised data' an elusive goal?

*Organised data is not in itself an elusive goal. However, when the volume and variety of data increase by orders or magnitude then maintaining organisation and coherence in the data is difficult to achieve and by extension makes the data difficult to use. Therefore, organised data - in the context of large heterogeneous datasets - is an elusive goal.*

Motivation:

p3: Would you not consider non-scientists looking at these images?

*In the article we use the term 'scientist' for anyone who aims to use data for some end. The claim is not that only scientists look at data; rather, anyone (formal scientist or not) who uses the data is referred to as a scientist. The order of terminology is important.*

p3: "in the use of 'ways and means' of effecting the organisation"  
-> I have no clue what this means.

p4: "To achieve this ... and so on"  
-> These vague statements need rephrasing (e.g. 'we define [...] to the \*various\* attributes'). Also, what is 'generally available equipment'?

*We have rephrased vague statements in line with this remark.*

p5: it is not entirely clear to me from reading the paper what the bandbox program does. The paper states that it is based on the 10 recommendations that follow, but as explained below the re-organisation in Figure 4 still violates several recommendations... Perhaps some pseudo-code may be useful? Also, wouldn't it make more sense to first describe the recommendations and then introduce this program?

*We have moved the description of what bandbox does to the Software Availability section.*

### **Recommendations:**

Except for recommendation (7), all recommendations start with the word 'consider'. Given these are recommendations, that is superfluous. It may be clearer to use an imperative to directly state the recommendation (like done in 7).

p6: How is a "raw" TIFF file defined?

*We have replaced this with the phrase 'uncompressed TIFF files'.*

(3a) "the fewer the better" means a depth of 1 is best. This is probably not what the authors intended.

*We accept the correction and have clarified the argument based on the ISA (investigation, study, assay) framework.*

(3b) Having a subfolder called 'tiff' is often a good idea, e.g. when there is also a file with metadata describing those tiff images (which is typically the case). In fact, the recommended Figure 4 has a 'raw' folder, which has exactly the same meaning, thus contradicting this recommendation.

*We accept the correction and have revised the text for clarity that we are referring to intermediate folders where none are required.*

(5a) What are "any references that are tied to the instrument" and why should they be

excluded? If these are references to the microscope used, they may be relevant to the user?

*We accept the correction and have revised the phrase.*

(5b) Why would dates in filenames be ambiguous and should they be avoided? Many data acquisition softwares write files with date and times in their names. Renaming these would, as the authors themselves point out, indeed be complicated and possibly lead to errors.

*The emphasis in the article is in having dates in folder names not file names. In (1b) we mention dates in file names as a possibility. Nevertheless, we do caution that date-time data on file names can also include subtle variations such as seconds so that numerous related files become non-trivial to work with due to these variations.*

(7a) I have no clue what this means: "similar folders at different depths have the same names"

*We have revised the recommendation and included an example with reference to Figure 3.*

(7b) What are "personal identifiers"?

*We have included a parenthetical remark with examples to illustrate what personal identifiers are.*

(7c) The name 'data' is actually used in the line below and in the recommended Figure 4. Also, I don't see why 'images' won't be an excellent name for a folder that contains images?

*These examples are purely for illustration purposes but are inspired by the actual structure used in EMPIAR in which the 'data' directory sits beside an XML file e.g.*

*[https://ftp.ebi.ac.uk/empiar/world\\_availability/10002/](https://ftp.ebi.ac.uk/empiar/world_availability/10002/), which we have omitted here. They were generated from the examples provided in the git repository.*

*We believe it is better to have descriptive folder names as opposed to generic names, which provide no meaningful information. The name 'images' does not convey any meaningful information. Better would be something like 'tomograms' or 'particles'. Nevertheless, this is configurable in bandbox using the bandbox/obvious\_files option in the configuration file.*

(7e) This recommendation may not be limited to slices of 3D data, which seems an arbitrarily narrow example for such broad recommendations. I personally thought of zero-padding images when I first read this (apparently not careful enough!). Perhaps using a term like "leading zeros" may be less ambiguous? Albeit perhaps useful to some of the readers, this is the only recommendation that has an explicit explanation of how to do this on two specific computer systems. Wouldn't this be something that could only be implemented in the bandbox program, so it could be used on any computer?

*We accept the correction and have rewritten the recommendation to clarify the context. We agree that implementing this in bandbox would enable a cross-platform solution and will plan this for a future release.*

(8) What are "friendly file formats?". Also, the term "widely used file formats" is not well defined.

*This has been revised to simply 'File formats'.*

(9a) The example in Figure 4 does not have a README file...

*A README file has been added in the updated figure.*

(9b) "This can be achieved ... data presents" -> This sounds superfluous and condescending.

*This sentence has been deleted in the article.*

p11: The proposed path "data/brief\_description/treatment3..." violates at least recommendations 3 (unused subfolder 'brief\_description') and 7 (use of word 'data')

*As mentioned above, the example used here is purely illustrative and omits other content which would otherwise not violate this recommendation.*

**Competing Interests:** No competing interests were disclosed.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact [research@f1000.com](mailto:research@f1000.com)

**F1000Research**