What we can learn from deep space communication for reproducible bioimaging and data analysis

Tatiana Woller[1,2], Christopher J. Cawthorne[3], Romain Raymond Agnes Slootmaekers[4], Ingrid Barcena Roig[5], Alexander Botzki[6], Sebastian Munck[7,2,#].

1. VIB Technology Training, Data Core, and VIB BioImaging Core, Ghent & Leuven, Belgium
2. KU Leuven, Neuroscience Département, Leuven, Belgium
3. Nuclear Medicine and Molecular Imaging, Department of Imaging and Pathology, KU Leuven, Leuven, Belgium
4. NSANGA, 36 Pakenstraat, Leuven, 3001, Belgium
5. Support for Research Data Management (RDM), KU Leuven, Leuven, Belgium
6. VIB Technology Training, Ghent, Belgium
7. VIB BioImaging Core, Leuven, Belgium

to whom correspondence should be addressed: Sebastian.munck@kuleuven.be

Abstract

Reproducible bioimage and data analysis is difficult. Multiple overlapping initiatives are currently being established, defining and recommending the annotation of images with metadata. However, good documentation of complex and evolving projects is difficult, and the currently fragmented landscape of electronic labnotebooks, metadata servers, repositories, and manuscripts is leading to either redundancy in annotation or omission, as well as a mix of inputs from different time points of the project. To tackle this, we take inspiration from space communication, where error correction protocols based on redundancy are built into data for transmission. Here, we provide a proof of concept using an Artificial Intelligence (AI) language model proofreading to digest the redundant metadata entries of this manuscript and visualize the differences between them to complete metadata entries, highlight inconsistencies, and correct human error to improve the overall documentation for more reproducibility and reusability.

Main

The reproducibility crisis describes the phenomenon that many scientific studies are difficult or impossible to reproduce, specifically affecting empirical studies. This irreproducibility, combined with pressure resulting from the fact that many research lines are publicly funded, threatens today's scientific endeavor's very fabric and credibility. A survey conducted by the journal Nature in 2016 showed that more than 70% of researchers attempted and did not succeed in reproducing someone else's experiments. In addition, more than half could not reproduce their own experiment (Baker, 2016). This lack of reproducibility is often attributed to multiple issues, most commonly a lack of access to raw data and its associated documentation, the inability to manage complex datasets, and the prevalence of studies with novel results over failed experiments (https://www.nature.com/articles/d42473-019-00004-y).

Even though reproducibility has been discussed widely in the last decade, it can be easily confounded with replicability, and its definition depends on the research domain. In biomedical research and computational biology (including Bioimaging), reproducibility indicates the ability to obtain the same results by using the same data and methods, while "replicability" stands for researchers arriving at the same scientific conclusion using their own data and methods (interestingly, the meanings of reproducibility and repeatability are swapped in computer science and microbiology (Plesser, 2018)). Both reproducibility and replicability can be split into different types. For instance, Goodman et al. make the distinction between methods reproducibility, results reproducibility, and inferential reproducibility. Methods reproducibility refers to providing enough metadata associated with data and the analysis to repeat the same analysis. By contrast, results reproducibility denotes the ability to

obtain the same results from a different student with a similar analysis as the original study. Finally, inferential reproducibility indicates drawing the same conclusion from a different study or reanalysis of the same study (Plesser, 2018). Machine learning and artificial intelligence are increasingly applied to bioimaging and are also confronted with reproducibility issues. Gunderson et al. define four types of reproducibility according to the documentation: R1 description, R2 code, R3 data, and R4 experiments (Gundersen & Kjensmo, 2018). The lowest degree of reproducibility is associated with R1 description because R1 description encompasses a textual description of the experiment. R2 code contains the code/workflow and its associated metadata but lacks the original data. R3 data refers to the available dataset and the associated metadata without the workflow for the metadata. R4 experiments indicate that the dataset, the code, and the associated documentation are available. R4 experiments denote the highest degree of reproducibility.

The FAIR smorgasbord

Approaching reproducibility through documentation is also a useful lens for viewing bioimaging data, and the multiple initiatives that have emerged within the bioimage analysts' world and beyond to promote reproducible research. These initiatives can be viewed as attempts to standardize the documentation that accompanies the generation of a bioimaging dataset (and here bioimage analysis can be seen as a placeholder for other data analysis disciplines). Organizations such as EOSC indicate the quality of the data and associated metadata will improve if it is FAIR, as in Findable, Accessible Interoperable, and Reusable (Wilkinson *et al*, 2016), which will enhance the reproducibility of research (https://zenodo.org/records/7515816). The FAIRification of the data requires work at many levels, ranging from ontology to reproducible analysis pipelines. Regarding ontologies for bioimage analysts, recent ontologies such as REMBI (Sarkans *et al*, 2021), MITI (Schapiro *et al*, 2022), and EDAM Bioimaging (https://bioportal.bioontology.org/ontologies/EDAM-BIOIMAGING) provide a starting point to report metadata associated with analysis. Several organizations such as NEUBIAS (https://eubias.org/NEUBIAS/), EuroBioimaging (https://www.eurobioimaging.eu/), and Elixir (https://elixir-europe.org/) promote the reproducible analysis by organizing training and developing tools. Among those tools, we can distinguish between workflow finders and benchmarking sites. Workflow finders enable tailoring a workflow that has been applied to a similar biological question. Common workflow finders are Biotools (https://bio.tools), Bioimage Model Zoo (https://bioimage.io/#/), and BioImage Informatics Index (https://biii.eu). Alternatively, to assess whether the workflow is the most suitable choice for a given problem, a benchmark comparison can be used such as BIAFLOWS (https://biaflows-sandbox.neubias.org/#/) or grand-Challenges (https://grand-challenge.org/challenges/).

However, when it comes to experimental documentation, the typical place for documentation is the electronic labnotebook (Myers *et al*, 2001). The aforementioned standardization of metadata like REMBI, MITI, EDAM, are somewhat distinct from the labnotebook. However, these metadata standards are increasingly incorporated with rich file formats (like OME-TIFF and OME -Zarr), data tools such as OMERO (that allow one to find and browse images alongside metadata; https://www.openmicroscopy.org/omero/) and ManGO (that allows associating any file with predetermined metadata on modern data management systems like iRODS (https://irods.org/; https://github.com/kuleuven/mango-metadata-schemas).

One of the consequences of using additional data annotations like REMBI, is that they can lead to a fragmentation of the metadata space: the coherence and consistency between labnotebooks, metadata servers, and a given manuscript is not always a given. Sample preparation data from the time of the experiment is typically found in the labnotebook, whilst acquisition and analysis information from an image is typically found in the associated image metadata file, so there is a

variable degree of redundancy when these sources are combined into a framework such as REMBI. However, as anyone who has had to "recover" data regarding an image from a labnotebook knows, redundancy can be a feature, not a bug. Seemingly cryptic information (time information, cell naming, filter idiosyncrasies, etc.) may allow confirmation of the identity of the related image file via probabilistic inference. This suggests that it can be valuable to keep electronic labnotebook entries, imaging metadata, and the final write-up as separate data entries, as they can all be used to reconstruct complete information in the presence of human error or "noise."

The use of redundant sources to deal with noise is well established in the field of space communication (itself of increasing interest given current international efforts for unmanned space exploration, e.g., returning to the moon (https://www.nbcnews.com/science/space/russia-india-china-us-are-heading-lunar-south-pole-rcna100495)). A central challenge that these probes must overcome is accurate communication. That information can be transmitted over noisy channels is based on Claude Shannon's landmark publication (Shannon, 1948). That it can be corrected is based on Richard Hamming's pioneering invention of the first error-correcting code in 1950 (Hamming, 1950). Today's telecommunication routinely employs different strategies for redundancy, error detection, and correction, including applications for unreliable storage mediums (e.g. EP2288991A2).

We can consider the creation of FAIR bioimage data as a similar challenge to that of deep space communication. It can be viewed as a transmission to the future (metaphorically a distant intelligent civilization) or to an entity with which direct communication is not possible (aka another lab, or a successor). As with space communication, the chance that information is lost or parts were not passed on properly and hence are unrecoverable, needs to be accounted for in the system.

Although in space communication the error correction code is intended to address noise that arises due to weak signal strength and distance, for bioimage data, we can consider the "noisy" channel as user documentation as above. In addition to the information included in publications (theoretically "stand-alone descriptions" containing all the information needed for replication); the lack of a universal/integrated system that leverages the electronic labnotebooks and the standardized metadata leads to both omission and (unused) redundancy.

In this commentary we propose to use the redundancy inherent in the different sources of data documentation (namely the electronic labnotebooks, metadata fileservers, manuscripts, GitHub resources, and images, etc. that comprise the typical 'data package' for a bioimaging experiment) to derive the most complete annotation and also enable cross correction if required, using a similar conceptual model. This should consolidate the different metadata sources, help to complete missing information, and overall future-proof the data and support the documentation for better reproducibility.

Application of AI

Whilst the call to proofread and consolidate various metadata entries might be noble, it is unlikely that many researchers will adopt it due to time constraints and its tedious nature. It is also understood that humans are notoriously bad at copy-editing. However artificial intelligence (AI) based language models have gained enormous popularity (with their rise having profound implications for (scientific) publications that will not be discussed here), are powerful for creating structured outputs, and can readily take over tedious proofreading tasks, suggesting they are ideal tools for complementing the human part of the documentation. Indeed commonly available language models like GPT-4 have been used for the post-hoc transformation of free-text radiology reports into structured reporting (Adams *et al*, 2023). Language models can consequently be used to

query if a specific diagnosis is present and to create a structured report for example in the form of a JSON file.

Analogous to this approach, for better reporting and improving reproducibility, the various metadata sources can be 'digested' by a state-of-the-art language model to generate a structured report file. Such a report can be based on the latest recommendations for metadata like, for example, REMBI or checklists as proposed by Schmied *et al*, (2023). Here gaps and contradictions can be highlighted: by using AI language model proofreading together with a metadata catalog, an error correction for the various metadata entries can be integrated and the overall annotation of the data improved. Here redundancy of different (meta)data sources and representation of their consistency can be considered similar to error correction in space and telecommunication as illustrated in **Figure 1A and B**. In addition, our analysis offers a feedback on the entries and their completeness to begin with.

Such a report could be published alongside a manuscript and could be a prerequisite for submission to a journal. It could be seen as analogous to a preregistered report, where the study proposal is peer-reviewed and bias is removed, only that here the documentation itself is retrospectively reviewed and reproducibility and long-term validity are enhanced.

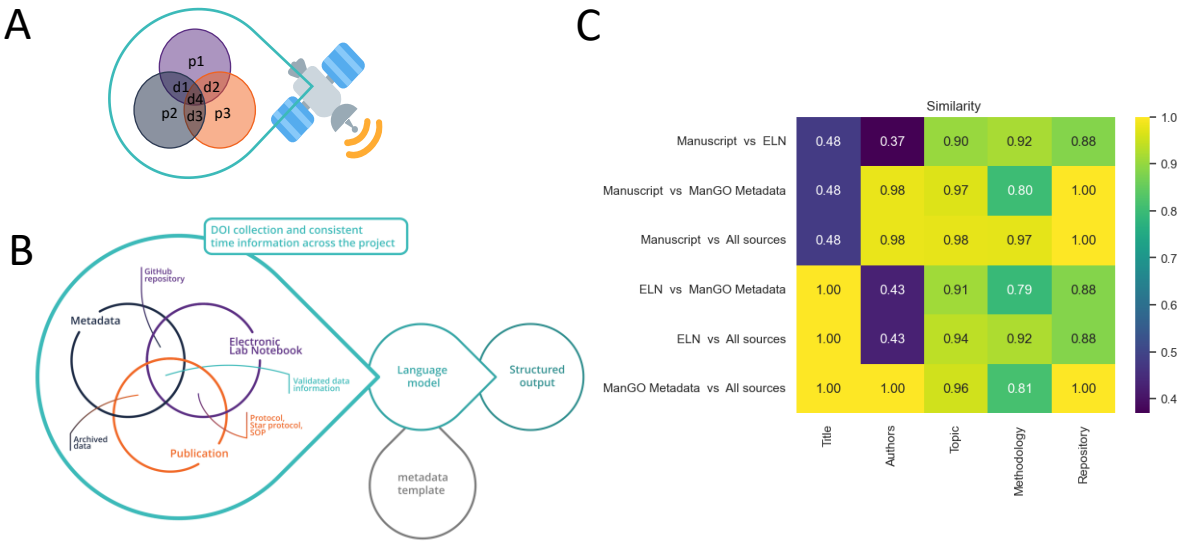

Figure 1. Linking metadata sources and digesting them with language models to generate structured outputs and representations of similarity. A) Illustration of Hamming code for error correction in data transmission. Transmission of data (d) and parity (p) bits allows for error correction via redundancy (https://en.wikipedia.org/wiki/Hamming_code). **B) Diagram showing the different sources of metadata information and how to bundle them.** Three independent resources: the electronic labnotebook, the data-associated metadata, and the publication are here shown as redundant entries. An AI language model can then be used to extract required and standardized data elements for verification, using codewords as a means of error correction analogous to error correction used in (space) communication. **C) Heatmap display for similarity between sources per keyword.** A Jupyter notebook using GPT-4 has been used to create a structured output in the form of a CSV file, (see supplementary Table 1). Here the digestion of a labnotebook entry, a metadata file server file, and this manuscript are used to check for keywords. The consistency of the keywords used across the sources is displayed in a heatmap using the cosine distance for semantic similarity estimation (https://spacy.io/api/doc).

Practical example

To create a proof of concept for the proposed approach we have created a workflow where the multimodal Large Language Model GPT-4 reads a labnotebook entry, a corresponding ManGO metadata file, and a publication. Using the labnotebook entry that started this project, we uploaded the image file of Figure 1B to the KU Leuven iRODS storage system, where we created a Metadata entry for the picture. Finally, we used this manuscript (sic!) with its content for testing consistency between entries. For simplicity we used a list of (5) keywords to check if the text that GPT-4 found related to these keywords were present and consistent across the (metadata) sources (with our list of five keywords we are effectively testing for the completeness of the title, the authors, the topic, the methodology, and the repository used; this list of keywords is a placeholder for a checklist, a metadata standard or a published template like REMBI as mentioned above). In the future aligning resources like REMBI or a recommendation list targeted at image analysis (Schmied *et al*, 2023) with the GPT-4 queries is desirable. We also see that this is a community effort, where metadata schemas are evolving and hopefully over time converging on a community agreed standard. Based on our five keyword query we created a Jupyter notebook (https://github.com/vib-bic-training/Reproducibility_RDM.git), which was used to create a report summarizing the findings of the proofreading and comparing the different entries (the exemplary labnotebook file used here as well as the metadata file can be found on the GitHub repository). We interacted with GPT4 using a API key. We also used a larger pipeline package (en_core_web_lg, 685k unique vectors) to carry out the tokenization, which could be customized towards a specific domain, such as BioImaging. The "digestion", the text that GPT-4 found in the manuscript, the labnotebook entry, and the metadata file based on the keywords are given in **supplementary Table 1**. It is impressive to see that even with these five simple keywords detailed descriptions can be extracted from the sources and compared. The consistency between the entries can be visualized in the form of a heatmap per keyword and source. **Figure 1C** shows a heatmap representing the similarity of the answers between the sources in **supplementary Table 1**. In a scenario where one of three data entries is different (e.g. a concentration), that value can now be corrected based on the majority of entries: the heatmap readily shows how similarity varies across the files (with a 1 describing perfect similarity).

Beyond the proof of concept stage and regarding the use of large language models in general, it is understood that the consistency over time, correctness, hallucinations, and confidence in the answer as well as the availability of the language model need to be monitored carefully for future implementations.

Overall, we believe that the application of the procedure outlined here can reduce errors in reporting and improve the overall reproducibility and FAIRness of bioimage data. With the possibility to provide easily interpretable readouts in the form of heatmaps that can highlight where metadata differs or is missing, we believe that information can be more easily completed and records consolidated. This should improve the overall quality of reporting and future-proof the reproducibility of the data and also improve the reusability of data as with good documentation trustability is enhanced for follow-up studies. Given the simplicity of the approach, it can be easily adopted, allowing image data to boldly go FAIR…(where too little data has gone before).

	Keywords				
Source	Title	Authors	Topic	Methodology	Repository
Manuscript	Proofreading Approach to Improve Reproducibility in BioImaging Metadata Using AI (Artificial Intelligence) Language Models	Tatiana Woller, Christopher J. Cawthorne, Romain Raymond, Agnes Slootmaekers, Ingrid Barcena Roig, Alex Botzki, Sebastian Munck	The paper discusses the challenge of maintaining reproducibility in BioImage data, similar to the challenges in deep space communications. It suggests the implementation of AI language model proofreading to improve error correction and consequently, the fidelity of the data.	The authors have created a workflow where the Large Language Model GPT-4 reads a lab notebook entry, a corresponding ManGO metadata file, and this publication to look for a list of keywords, and checks them for consistency across these metadata sources. Error correction is then performed by highlighting gaps and contradictions.	https://github.com/vib-bic-training/Reproducibility_RDM.git
ELN	What we can learn from deep space communication for reproducible BioImaging and data analysis	Not mentioned in the extract provided.	The project deals with reproducible BioImaging and data analysis, adopting techniques from deep space communication. It involves running scripts interactively with GPT-4 through an API.	The scripts for this project are executed using Python 3.10. Dependencies are listed in the requirements_sm.txt file. The main interaction with GPT-4 is done via an API for which an API key is required from OpenAI. The accepted input formats for the scripts are pdf, txt, and json files.	The Github repository for the scripts and config files is found at https://github.com/vib-bic-training/Reproducibility_RDM. The project is based on https://github.com/kbressem/gpt4-structured-reporting.
ManGO Metadata	What we can learn from deep space communication for reproducible BioImaging and data analysis	Tatiana Woller, Christopher Cawthorne, Romain Raymond Agnes Slootmaekers, Ingrid Barcena Roig, Alex Botzki, Sebastian Munck	The study discusses learning from deep space communication to improve the reproducibility of BioImaging and data analysis.	The research uses a Large Language Model (GPT-4) and follows the REMBI standard for their analyses.	The code and data used for the study can be found at https://github.com/vib-bic-training/Reproducibility_RDM.git.
All sources	What we can learn from deep space communication for reproducible BioImaging and data analysis	Tatiana Woller, Christopher Cawthorne, Romain Raymond Agnes Slootmaekers, Ingrid Barcena Roig, Alex Botzki, Sebastian Munck	This paper discusses the issues regarding reproducibility in BioImage and data analysis. It offers a solution using the concept of error correction protocols used in space communication. The authors propose the use of AI language model proofreading to digest redundant metadata entries, visualize the differences, and correct errors. This method aims to increase metadata consistency, and improve overall documentation for higher reproducibility and reusability.	The authors implemented a workflow, where the GPT-4 Language Model reads different metadata sources, including a lab notebook entry, the corresponding ManGO metadata file, and this publication. These sources are searched for specific keywords to see their consistency. Then, conflicts and discrepancies are highlighted, allowing for error correction and improvement of the fidelity of the annotation.	https://github.com/vib-bic-training/Reproducibility_RDM.git

Supplementary Table1. Proof of concept report. A Jupyter notebook using GPT-4 has been used to create a structured output in the form of a Table (CSV file). The digestion of a labnotebook entry, a ManGO metadata file and this manuscript are used to check for keywords and their consistent use.

Acknowledgements

The authors like to thank Christof De Bo for help with the Figure design. Part of the resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. SM an CC are supported by FWO I000123N Flanders BioImaging: Leading Imaging Application Integrated Service and Enablement (FBI-LIAISE). SM is supported by FWO I001322N - 3D SUper-Resolution to cryo-Electron MIcroscopy to study nanoscale subcellular dynamics and structure that alter in Neurodegenerative Diseases - 3SURE MIND.

References

Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR & Bressem KK (2023) Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. *Radiology* 307: e230725

Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature* 533: 452–454

Gundersen OE & Kjensmo S (2018) State of the Art: Reproducibility in Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence* 32

Hamming RW (1950) Error detecting and error correcting codes. *The Bell System Technical Journal* 29: 147–160

Myers J, Mendoza E & Hoopes B (2001) A Collaborative Electronic Laboratory Notebook. (https://papers.ssrn.com/abstract=2969589) [PREPRINT]

Plesser HE (2018) Reproducibility vs. Replicability: A Brief History of a Confused Terminology. *Frontiers in Neuroinformatics* 11

Sarkans U, Chiu W, Collinson L, Darrow MC, Ellenberg J, Grunwald D, Hériché J-K, Iudin A, Martins GG, Meehan T, *et al* (2021) REMBI: Recommended Metadata for Biological Images—enabling reuse of microscopy data in biology. *Nat Methods* 18: 1418–1422

Schapiro D, Yapp C, Sokolov A, Reynolds SM, Chen Y-A, Sudar D, Xie Y, Muhlich J, Arias-Camison R, Arena S, *et al* (2022) MITI minimum information guidelines for highly multiplexed tissue images. *Nat Methods* 19: 262–267

Schmied C, Nelson MS, Avilov S, Bakker G-J, Bertocchi C, Bischof J, Boehm U, Brocher J, Carvalho MT, Chiritescu C, *et al* (2023) Community-developed checklists for publishing images and image analyses. *Nat Methods*: 1–12

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423

Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, *et al* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3: 160018