What we can learn from deep space communication for reproducible BioImage analysis

Tatiana Woller[1,2], Chris 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, Belgium
2. KU Leuven, Neuroscience Department, 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

Space communication requires error correction protocols based on redundancy built into data for transmission. As data transmission is also central to reproducible BioImage analysis (with noise often resulting from ill-defined annotation), we advocate here for error correction based on redundant documentation, combined with AI language model proofreading.

Main

Recent news highlights increased international efforts for unmanned space exploration, e.g. returning to the moon, (Space race 2.0: Russia, India, China and the U.S. are heading for the lunar south pole, 2023). 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 knows different strategies for redundancy, error detection, and correction including applications for unreliable storage mediums (Spiegeleer & Slootmaekers, 2011).

Creating FAIR BioImage data can be considered a similar challenge to that of deep space communications: it can be viewed as 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): just as with space communication the chance that information is lost or that 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 could consider the "noisy" channel as user-entered data. Whilst efforts have been made towards improving experimental documentation via electronic lab notebooks (Myers *et al*, 2001), standardizing metadata (Sarkans *et al*, 2021; Norgaard *et al*, 2022) with rich file formats (OME -Zarr) (Moore *et al*, 2023), creating data tools such as OMERO (that allow one to find and browse images alongside metadata) (Allan *et al*, 2012) and ManGO (that allows associating any file with predetermined metadata on modern data management systems like iRODS (Conway *et al*, 2011) ;https://github.com/kuleuven/mango-metadata-schemas (Ghosh *et al*, 2023)) 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 leads to both omission and (unused) redundancy.

As anyone who has had to “recover” data regarding an image from a lab notebook knows however, 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 lab notebook 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”: in effect serving an analogous function to the requirement for redundancy for forward error correction used in space communication (but only if the information can successfully be extracted by the user or receiver with an equivalent to the parity bits/checksum code).

Recently, 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). These tools 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 documentation. Indeed, in a recent and relevant publication, 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.

Here we propose to use a similar approach to proofread different sources of metadata, namely the electronic labnotebooks, metadata fileservers, manuscripts, GitHub resources and images etc. that comprise the typical ‘data package’ for a bioimaging experiment. These various 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 (Sarkans *et al*, 2021) or checklists as proposed by Schmied *et al*, (2023). Here, importantly, gaps and contradictions can be highlighted. In this sense, by AI language model proofreading together with a metadata catalog, an error correction for the various metadata entries can be integrated and the overall fidelity and trustability of the data improved. **Figure 1** links these different (meta)data sources in a representation that is akin to visualizations used for error correction as used for space communication.

Technically, such a report could be published alongside a manuscript and could even be a prerequisite for submission to a journal or peer review.

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. 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).

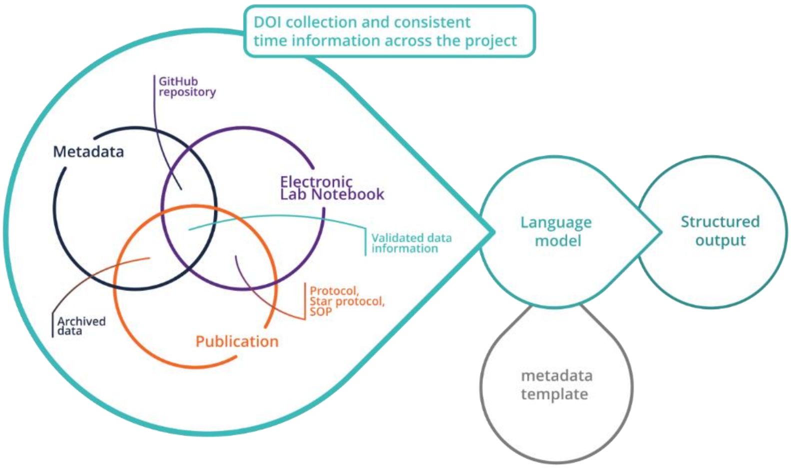

Figure 1. Diagram showing the different sources of metadata information and how to bundle them. Three independent resources: the electronic lab notebook, the data-associated metadata, and the publication are here shown as redundant entries. An AI language model can then be used to extract required and standardized data elements for verification, using codewords as a means of error correction analogous to error correction used in (space) communication.

Materials and Methods

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 this publication to look for a list of (5) keywords to check if they are present and consistent across the (metadata) sources (see **Supplementary Table 1**). The list of keywords is a placeholder for a checklist, a metadata standard or a published template as mentioned above. The created Jupyter notebook (https://github.com/vib-bic-training/Reproducibility_RDM.git) is then used to create a report summarizing the findings of the proofreading and comparing the different entries and correct errors given enough redundancy. The exemplary labnotebook file used here as well as the meta-data file can be found on the GitHub repository as well. The consistency between the entries can be further visualized in the form of a heatmap per keyword. **Figure 2** show an example for the Title. Beyond the proof of concept stage It is understood that the consistency over time, the correctness, hallucinations, as well confidence of the answer and the availability of the language model needs to be monitored carefully for future implementations and other language models may apply.

	keywords				
Sources	Title	Authors	Topic	methodology	repository
Manuscript	"Proofreading Approach to Improve Reproducibility in BioImaging Metadata Using AI (Artificial Intelligence) Language Models "	Tatiana Woller, Chris Cawthorne, Romain Raymond Agnes Slootmaekers, Ingrid Barcena Roig, Alex Botzki, and Sebastian Munck	The paper focuses on the challenge of maintaining the integrity and reproducibility of BioImage data, similar to deep space communications. It proposes a solution through AI language model proofreading to enhance error correction and improve data fidelity	The authors developed a workflow using the GPT-4 Large Language Model to proofread different sources of metadata. This involved the model reading a lab notebook entry, a ManGO metadata file, and the paper itself to check for consistency of keywords across these sources. Identified gaps and contradictions could then be highlighted for error correction. This approach was implemented as a proof of concept using a Jupyter notebook	https://github.com/vib-bic-training/Reproducibility_RDM/tree/main/examples
ELN	"What we can learn from deep space communication for reproducible BioImage analysis"	The authors are not mentioned in the extract provided.	The main topic of this publication is exploring how deep space communication principles can be implemented for reproducible BioImage analysis, using the GPT-4 API from OpenAI	The methodology involves obtaining the scripts and config files from the linked GitHub repository. After installation, an analysis is run through the interaction with the GPT-4 API. The accepted input formats for the analysis are pdf, txt, and json files. The work, however, is limited to GPT-4 from OpenAI	https://github.com/vib-bic-training/Reproducibility_RDM/tree/main/examples
Mango metadata	"What we can learn from deep space communication for reproducible BioImage analysis"	Tatiana Woller, Christopher Cawthorne, Romain Raymond Agnes Slootmaekers4, Ingrid Barcena Roig, Alex Botzki, Sebastian Munck	Deep space communication and its learnings for reproducible BioImage analysis	Large language model (GPT-4)	https://github.com/vib-bic-training/Reproducibility_RDM.git
All sources	"What we can learn from deep space communication for reproducible BioImage analysis"	Tatiana Woller - Christopher Cawthorne - Romain Raymond Agnes Slootmaekers - Ingrid Barcena Roig - Alex Botzki - Sebastian Munck	Applying error correction principles from deep space communication to increase the reproducibility of BioImage analysis.	Utilizing AI language models (specifically GPT-4) to proofread different sources of metadata, namely the electronic lab notebooks, metadata file servers, manuscripts, and GitHub resources that comprise a typical 'data package for bioimaging experiment. These sources are, then, digested by a state of the art language model to generate a structured report file.	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.

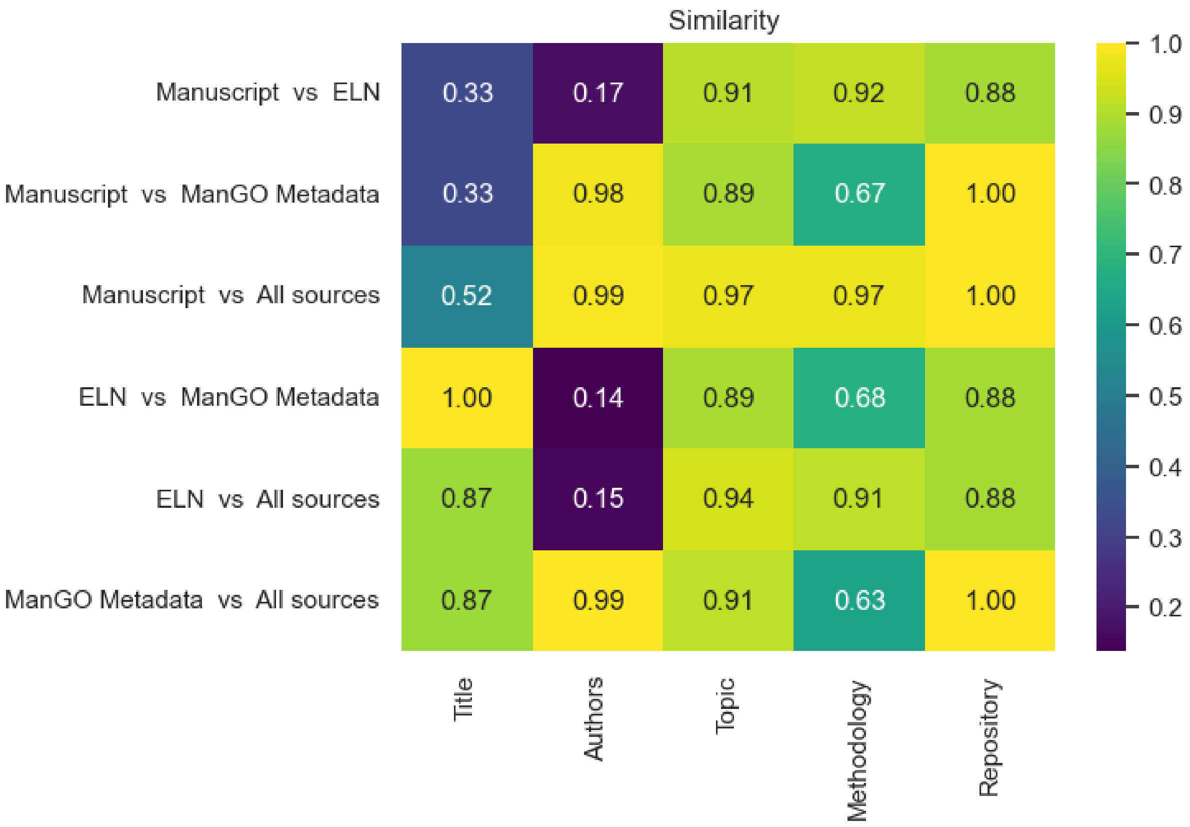

Figure 2. Heatmap display for similarity between sources per keyword. A Jupyter notebook using GPT-4 has been used to create a structured output in the form of a CSV file, (see supplementary Table 1). Here the digestion of a labnotebook entry, a metadata file server file and this manuscript are used 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 (Doc · spaCy API Documentation).

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

Allan C, Burel J-M, Moore J, Blackburn C, Linkert M, Loynton S, MacDonald D, Moore WJ, Neves C, Patterson A, *et al* (2012) OMERO: flexible, model-driven data management for experimental biology. *Nat Methods* 9: 245–253

Conway M, Moore R, Rajasekar A & Nief J-Y (2011) Demonstration of Policy-Guided Data Preservation Using iRODS. In *2011 IEEE International Symposium on Policies for Distributed Systems and Networks* pp 173–174.

Doc · spaCy API Documentation *Doc*

Ghosh M, Broothaerts K, Ronsmans S, Roig IB, Scheepers J, Dikmen M, Ciscato ER, Blanch C, Plusquin M, Nygaard UC, *et al* (2023) Data management and protection in occupational and environmental exposome research - A case study from the EU-funded EXIMIOUS project. *Environmental Research* 237: 116886

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

Moore J, Basurto-Lozada D, Besson S, Bogovic J, Bragantini J, Brown EM, Burel J-M, Casas Moreno X, de Medeiros G, Diel EE, *et al* (2023) OME-Zarr: a cloud-optimized bioimaging file format with international community support. *Histochem Cell Biol* 160: 223–251

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

Norgaard M, Matheson GJ, Hansen HD, Thomas A, Searle G, Rizzo G, Veronese M, Giacomel A, Yaqub M, Tonietto M, *et al* (2022) PET-BIDS, an extension to the brain imaging data structure for positron emission tomography. *Sci Data* 9: 65

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

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

Space race 2.0: Russia, India, China and the U.S. are heading for the lunar south pole (2023) *NBC News*

Spiegeleer KD & Slootmaekers RRA (2011) Method of storing a data set in a distributed storage system, distributed storage system and computer program product for use with said method.