TITLE PAGE

Citation Format:

Vamshi Damagatla, Alessandro Bossi, Ilaria Bargigia, Antonio Pifferi, "A Data Management Plan for Open Data, readable and reusable by humans and AI", Proc. SPIE 13935, Diffuse Optical Spectroscopy and Imaging X, 139350W (December 18, 2025). DOI: https://doi.org/10.1117/12.3098409

Abstract link:

https://www.spiedigitallibrary.org/conference-proceedings-of-spie/13935/139352W/A-data-management-plan-for-open-data-readable-and-reusable/10.1117/12.3098409.short

Copyright notice:

A Data Management Plan for Open Data, readable and reusable by humans and AI

Vamshi Damagatla,[1,*] Alessandro Bossi,[1] Ilaria Bargigia,[1,2] Antonio Pifferi,[1,3]

[1] *Politecnico di Milano, Dipartimento di Fisica, Milan, 20133, Italy*
[2] *Center for Nano Science and Technology, Istituto Italiano di Tecnologia@PoliMI, Milan, 20134, Italy*
[3] *Istituto di Fotonica e Nanotecnologie, Consiglio Nazionale delle Ricerche, 20133, Milan, Italy*

[*] *saivamshi.damagatla@polimi.it*

Abstract: We present a general and comprehensive data management plan to create open data which is both human and machine readable. We test and demonstrate its versatility using open AI engines to read and reuse it.

1. Introduction

In today's scientific world, the importance of data cannot be undermined as it forms the foundation of most scientific research. The 'Open Science' movement has been revolutionizing the research community and is important going towards future endeavours. While it contains many definitions and parameters, the three common practical aspects are - open access, open source and open data. While the first two have been underway with the shift towards open access publishing, and availability of open source codes and repositories, the final point of open data indeed poses a huge challenge, as it opens a Pandora's box of the looming question, 'What counts as data?'. While many regulations and frameworks are being formulated [1], they have not yet been concretized. However, despite all entities and governing bodies, the most important crux of open data is every single input researcher who will be the 'creator' of that data. And to go from multiple, innumerable creators to one big connected tree of data requires a meticulous planning so that all these data can co-exist, harmonize and interact with each other. At the same time, any part of this big tree of data should be understandable enough to any general 'user' of that data. To do this, the most imperative need is the formulation of a Data Management Plan (DMP).

2. 'Data Management Plan' (DMP) for Open Data

The formulation of a good DMP is of utmost importance as it lays framework to be followed in the long term. In our research group, we work in the field of time-domain diffuse optical spectroscopy (TD-DOS), and it was our aim to draft a DMP, that at its lower levels has a generality to encompass the different research performed across various laboratories, and yet at its higher, individual specific levels, allows each creator to fit it to their particular need. Hence, we divided the dataset into four different sections as shown in Fig. 1.

1. Overview - A brief verbose run-through of the dataset and the various files it contains. While it can be considered as a metadata to understand the dataset, it provides an idea and a roadmap to a potential user to understand what the dataset contains, and help them understand how to navigate and use it.

2. Data - The actual data, divided into *'Raw data'* and *'Meta data'*. While *'Raw data'* can be the direct output of the scientific instruments ranging from surveys, outputs of spectrometers, oscilloscopes, MRI machines, or time-of-flight curves in our case *'Meta data'* consists of supporting information about the particular experiment used to acquire the data that could be manually written, photographed, etc such as samples, subjects, wavelengths, time bins, etc in our case.

3. Tools - The requirements necessary to visualize, read or analyse the raw data. These could range from software codes to read or visualize encrypted data, to algorithms to analyse them, or even manuals of instruments used, etc.

4. Results - The final section is perhaps the tricky part and contains output results obtained and could range from preliminary results to published results in case of scientific publications.

Diffuse Optical Spectroscopy and Imaging X, edited by Davide Contini, Yoko Hoshi, Thomas D. O'Sullivan, Proc. of SPIE-Optica Vol. 13935, 139352W · 2025
Published by SPIE-Optica · 1605-7422 · doi: 10.1117/12.3098409

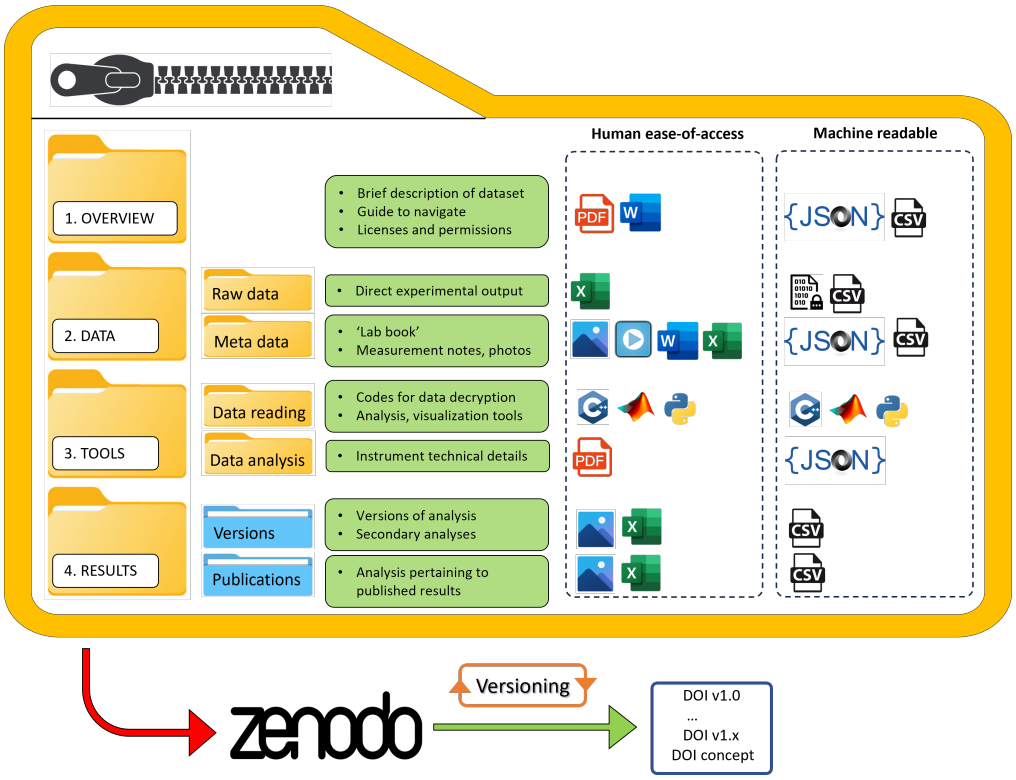

Fig. 1. Schematic of the data management plan (DMP)

3. Making the DMP FAIR

The final aim of open data is to not of just data dumping, but rather to make it abide by the FAIR principles - **F**indable, **A**ccessible, **I**nteroperable and **R**eusable [2]. With these in mind, we incorporated some helpful, yet not restrictive guidelines into our DMP. To aid with the interoperability, we have to ensure that it is easily understandable to a human user, while at the same time being machine readable. As shown in Fig. 1, we ensured this by using various formats to help both human and machine users access and use the data. To ensure it is findable and reusable, we had decided to proceed with the repository ZENODO as our preferred choice for data storage. While its being managed by CERN on behalf of the European Commission makes it a trusted and well managed, it offers certain other benefits. Its policy of providing a citable digital object identifier (DOI) makes it also findable and reusable. The option to reserve a DOI while in the draft stage throws up possibilities of referencing said DOIs in scientific articles, pre-publication, thus linking published research with the underlying data and making it findable in the process. With the availability of data versioning, while data cannot be deleted, new versions can be updated thus enabling a systematic building up on existing datasets. The continuous development of the platform, and the ability to cross link data hints to the direction of creation of truly open datasets across times and disciplines. Further, while the raw data in itself might be in encrypted or in proprietary formats, the meta data pertaining to them along with the tools to operate on them must utilize preferably open source or widely used resources. An example of this could be .TXT or MS-OFFICE for metadata and PYTHON or MATLAB for software codes.

4. Dataset creation in practice

While the initial DMP took time to be created, discussed, modified according to various inputs across research teams, the implementation also played a huge role. Every new dataset created helped in further refinements of the DMP. It is important to note that DMP should be a live document and should be constantly updated to adapt to new challenges to accommodate as many creators and users as possible. Indeed, having first tested it with researchers who were involved in the creation of the DMP, we then tested its versatility on other research teams in similar fields and received a good response for its use and adaptability. It was then opened to a larger group of creators from various labs through the organization of DATATHONs and this is currently an exercise in progress. Using this template, we were able to create more than 10 comprehensive datasets which have been since deployed

online. Further, this has also led to the the compilation of a comprehensive dataset on *in-vivo*, broadband, TD-DOS measurements performed on volunteers [3] and is ready to be published as an open data paper. This dataset contains the optical spectra of 10 subjects at 5 different locations, and reinforces our idea of open data not only being available for the sake of transparency and review, but also to creating a comprehensive dataset of measurements for others who might not have access to our instrumentation or techniques, so that they can join the community and aid in joint research in the spirit of open science.

5. AI readability and re-usability

While researchers were able to read the datasets with ease, the final challenge was to test its machine readability. Hence, we gave one of our datasets as an input to an open AI platform - ChatGPT-4o [4], and indeed the AI was able to read our data and output specific data as asked, for example - a DTOF curve for subject 6, on the forearm, at 700 nm. Having verified thus other deployed datasets, we can say that our DMP does help create truly machine

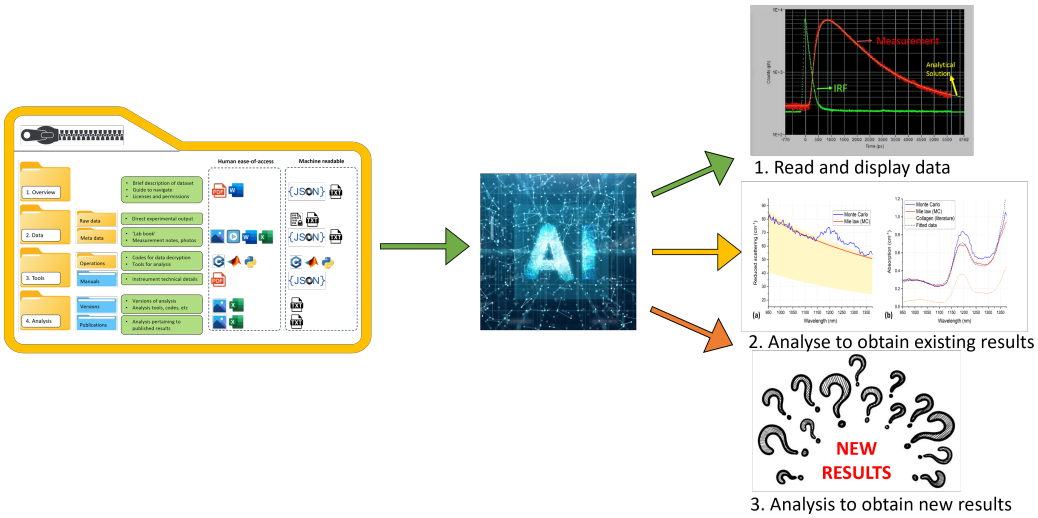

Fig. 2. Possible applications of AI to read and reuse open datasets

readable datasets. This is useful to potential users who wish to visualize the data before undertaking further work on it. Further, it also eases the process of data interoperability. This was indeed a promising and exciting result that can be demonstrated with ease in any AI model at any time. Taking it to the next step, while the AI can read our data, we wished now to check if AI can even analyse our data using our analysis methodology included as part of the dataset, to obtain back the results we had obtained in our work. This can be useful to peer review research work being done or help new users obtain our results as a verification of our experiments. Work is still ongoing to see if it can be done using our methodologies and even using commands for general analysis techniques based on the theoretical information found online. A last and final step for the future could be to see if we could use AI to analyse and yield new results, outcomes or perspectives from our existing data. This could be hugely useful in analysing huge amounts of data and finding patterns and correlations across datasets spanning larger time periods, and this cannot be possible without a generally usable DMP.

Acknowledgments

The authors acknowledge funding from (i) Horizon2020 PHAST-ETN (ID:860185); (ii) NextGeneration EU I-PHOQS (R0000016, ID D2B8D520); (iii) Next Generation EU-“PNRR-M4C2, “PRIN 2022 fund” (ID:20225MR35K);

References

1. OpenAIRE. https://www.openaire.eu/.
2. M. Axton, et al. ”The FAIR Guiding Principles for scientific data management and stewardship.” Sci. data 3.1 (2016).
3. I. Bargigia, et al. ”Comp. dataset of abs. and scat. spec. of in-vivo biological tissues using time-domain diffuse optical spectroscopy.” doi.org/10.1364/OTS.2024.OS3D.8
4. J. Achiam, et al. ”Gpt-4 technical report.” arXiv preprint arXiv:2303.08774 (2023).