LIS 545 B Wi 23: Data Curation I: Fundamentals
Term Project - Final
By: Javier Martinez Rodriguez [jjmr92]

Data and Metadata Profile

Dataset:
https://figshare.com/articles/dataset/_Does_the_Brain_Detect_3G_Mobile_Phone_Radiation_Peaks_An
_Explorative_In_Depth_Analysis_of_an_Experimental_Study_/1410034?file=6752208

1. Data

These data attempt to analyze brain activity when exposed to 3G cell phone signals. Each data point represents an EEG reading from female participants who were exposed to 3G cell phone signals as well as two sham cell phone conditions. These data come from a study conducted at Maastricht University sometime before 2015 (the actual times of data collection are not clearly stated so the timeline is inferred based on the date of publication for the results).

The main stakeholders for these data are possibly the authors or creators who have analyzed it in the experiment they conducted. Another possible stakeholder would be readers of the published results who may be interested in auditing the data. In the same vein, a third key stakeholder would be other scientists who may be seeking to replicate the data analysis or the experiment to validate or audit previous results. Sponsors of the research project may also be interested in these data as they may make it available through their own repositories.

There are three data files in three different formats: .SAV, .DOCX, and .SPS. The first file ('S1 Datafile.SAV') contains the raw data that was collected. The second file ('S1 Text.DOCX') is an example of SPSS syntax for multilevel regression analysis. The third file ('S2 Text.SPS') contains SPSS syntax that was used for analyzing the data and is available so users of the data can replicate the author's analysis.

This dataset has no usage restrictions. According to the licensing information provided in the dataset record, this data can be shared and reproduced for any purposes including commercial use. Although the record does not contain any information about specific software requirements for usage of these data, inferences can be made based on the file formatting. Microsoft Word is the recommended software for working with .docx files. The remaining two files can be best used with SPSS which is a statistical tool developed by IBM. There are licensing costs associated with the use of SPSS which is a major consideration for the use of these data.

2. Metadata

This dataset contains some metadata in the sense that there is a header row in the file containing all the data ('S1 Datafile.SAV'). This header row seems to contain information about the data in the columns; however, this header row contains information in Dutch and there is no further information elaborating on what the metadata for each column means or what the criteria is for data to go into a certain column.

For the dataset as a whole, there is plenty of information that has been provided through the repository. The repository record contains a title, a description of the dataset, the names of the authors, the name of the publisher, and the dates in which the dataset was posted to the repository as well as the first online date. Furthermore, this record contains keywords which are relevant for the specific research project and category information which serve as subject headings for this entire repository. The metadata also includes licensing information for this dataset and a link to the research article produced from the analysis of these data.

My perception is that the metadata for this dataset is fairly comprehensive, there is enough information for it to be found through many different channels. However, I believe the metadata within the actual data file is insufficient in the sense that this file lacks the necessary documentation for the understanding of what this metadata means or what its purpose would be. It is possible that someone working in this field would be able to understand it without further documentation. It is not possible for me to determine whether this metadata follows any particular standard. For the repository, it does seem like some standard metadata has been required from creators of datasets with some of it being mandatory.

3. Metadata Enrichment

The metadata record within this repository seems complete enough for this particular file. If I had to include any further information, I would include more detailed metadata for each file. I believe a description of each file would be sufficient for providing additional context to the dataset. With regards to the actual data, I believe that additional documentation would be required for the metadata within the data file. I would include more information about the metadata, what is being used and for what purposes as well as documentation on naming conventions. I believe this would allow those that are unfamiliar with the data to use them without having to guess or decipher their elements.

4. Publications

One research article has been published with the use of this data. The title of this article is "Does the Brain Detect 3G Mobile Phone Radiation Peaks? An Explorative In-Depth Analysis of an Experimental Study." This dataset shares a title with this article which is linked to the dataset record. I was not capable of finding, in a simple search, any further literature that cites these data as a source for information. Most of the information that can be found about the literature related to this dataset is actually citing the results in the aforementioned research article. My guess would be that sharing a title with the research article makes it more complicated to find publications that quote these data in particular. I performed my search using the title as I found the titles of the files to be non-specific and unlikely to yield better results.

Article:

Roggeveen, Suzanne; van Os, Jim; Lousberg, Richel (2016): Does the Brain Detect 3G Mobile Phone Radiation Peaks? An Explorative In-Depth Analysis of an Experimental Study. PLOS ONE. Dataset. https://doi.org/10.1371/journal.pone.0125390

Repository Profile

Repository URL: https://openneuro.org/
Metadata Standard: https://bids.neuroimaging.io/

1. Why did you choose this repository?
I chose to use OpenNeuro as a repository for a couple reasons: (1) it is a repository dedicated to neuroimaging data, (2) it is used for the sharing of academic data, and (3) it offers open access to the data within the repository and does not accept data used for commercial purposes.

First, the data set that I have chosen to work with includes EEG data from participants in an experiment. Because of this, in my search for a repository, I focused on those with a more limited scope. My intention is to include these data in a repository that users seeking neuroimaging data would consult. OpenNeuro stores neuroimaging data from different neuroimaging modalities which makes it a comprehensive repository for neuroscientists. Second, the repository is used exclusively for data tied to academic research. One of the considerations for uploading data to this repository is that it must be publicly available within 36 months of the first successful version of the dataset (Stanford Center for Reproducible Neuroscience, n.d.-a) which guarantees the data will support academic research. Finally, the repository ensures the data is publicly available with public domain licensing which matches the dataset I selected.

2. Is the repository open for data submissions from anybody, or does it have a defined collection scope?
OpenNeuro is in a gray area as it pertains to this question. It is technically open for data submissions from anybody, in the sense that anybody can create an account and submit data. However, the scope of the data allowed in the repository is limited.

3. Closed to general submissions,
3.1 Collection scope and eligibility
The collection is limited to neuroimaging data that is also made publicly available and BIDS-compliant. Although this is not made explicit by any of the information about the repository, I would think it safe to assume that the data comes from experiments in academic settings which are likely to be grant funded. This is an assumption made based on anecdotal experience. The repository itself does not place any

limits related to the origins of the data beyond it being ethically obtained and unidentifiable (Stanford Center for Reproducible Neuroscience, n.d.-a).

4. Open for submissions,

4.1. Acceptable data and file formats

The repository will accept neuroimaging data from the following types of neuroimaging modalities: MRI, PET, MEG, EEG, and iEEG (Stanford Center for Reproducible Neuroscience, n.d.-b). The only restrictions related to the actual data is that it must be BIDS-compliant. Considering the nature of the science, the data stored here is most likely to be experimental or observational. It is part of the medical science field, specifically neuroscience. As stated before, the repository only accepts data in what they call "BIDS format." BIDS (Brain Imaging Data Structure) is a metadata standard that has been developed for the organization and description of data within the field of Neuroscience (Brain Imaging Data Structure, 2023). BIDS itself requires the data to be in two formats: .json and .tsv. (Brain Imaging Data Structure, 2021b). The section of BIDS' website on Metadata outlines the programs that can be used to work with these formats such as: Python, R, Matlab, among others.

4.2. Submission Information Package (SIP)

The only provided information for what should be in an SIP is that the data must be BIDS-compliant. After the submission of a BIDS-compliant file, the file will go through BIDS validation and, if the validation does not generate any errors, the file will be uploaded to the repository. Following this, the user is prompted to fill out a metadata form and accept the terms and conditions of the repository (Stanford Center for Reproducible Neuroscience, 2022). More specifically, for a dataset to be BIDS-compliant it should be stored in folders organized hierarchically and in .json, .tsv, and any raw data format (Brain Imaging Data Structure, 2021a). A BIDS-compliant dataset must contain four folders in the following order in a top-down hierarchy: project, subject, session, and datatype (Brain Imaging Data Structure, 2021a). It is then necessary for a SIP to include all four of these folders so the data can be uploaded to OpenNeuro.

4.3. Human assistance or consulting

OpenNeuro, in its home page, includes a search bar for the articles in its documentation as well as an option to obtain human assistance via email by filling out a short form outlining any issues or questions the user may have. Users may also email their help desk directly.

4.4. Metadata and standards

As outlined above, the metadata included in OpenNeuro is submitted via a form that users fill out after uploading their file. However, as mentioned, the file itself must follow the BIDS standards which require users to provide metadata in .json format using a "key": "value" structure. Additionally, data tables must be submitted in .tsv format with the metadata for the content of each column being labeled in the first row of the data table (Brain Imaging Data Structure, 2021b).

5. Data Access Mechanisms

5.1. Is a login required?

A login is not required in order to access or download any data. A login is only required for the submission of data.

5.2. Types of data access mechanisms

OpenNeuro allows users to download data in a number of different ways: (1) data may be directly downloaded with a browser to a local file, (2) download from S3, (3) download with node.js, or (4) download from DataLad or GitAnnex from GitHub. Information and instructions on how to use any of these access mechanisms are available in any of the records for each dataset in the repository.

5.3. Does the repository display metadata using any specific metadata standard?

Each dataset in OpenNeuro has metadata available for user access through the metadata tab. The metadata is displayed according to the standards created by the repository developers. I am unable to discern whether or not this follows any other widely adopted standard since the information is mostly applicable to neuroscience.

5.4. Dissemination Information Package (DIP)

The DIP for the records in this repository include: a dataset description from the README file in the dataset, the data organized according to BIDS, an accession number, authors of the dataset, modalities for data collection (e.g. EEG, MRI, etc.), version codes, a description of the task for the data collection, name of the person who uploaded the file, date that the file was uploaded, number of sessions, number of participants, a DOI, licensing information, citation information, acknowledgements, funding information, references to articles supported by the data, information on ethics approval.

References

Brain Imaging Data Structure. (2021a). *Folders—BIDS starter kit*. Folders.

https://bids-standard.github.io/bids-starter-kit/folders_and_files/folders.html

Brain Imaging Data Structure. (2021b). *Metadata and file formats—BIDS starter kit*. Metadata and File

Formats. https://bids-standard.github.io/bids-starter-kit/folders_and_files/metadata.html

Brain Imaging Data Structure. (2023, February 9). *About*. About - Brain Imaging Data Structure.

https://bids.neuroimaging.io/index

Stanford Center for Reproducible Neuroscience. (n.d.-a). *Frequently Asked Questions—OpenNeuro*.

Frequently Asked Questions - OpenNeuro. Retrieved February 18, 2023, from

https://openneuro.org/faq

Stanford Center for Reproducible Neuroscience. (n.d.-b). *OpenNeuro*. Homepage. Retrieved February 18,

2023, from https://openneuro.org/

Stanford Center for Reproducible Neuroscience. (2022). *User Guide—OpenNeuro documentation*. User

Guide. https://docs.openneuro.org/user_guide.html

Final Considerations

1. Recommended Data Citation

There is no current citation for this specific dataset. The citation that is generated by figshare directs users to the journal article that was created based on the analysis of these data. My recommendation would be to create a permanent identifier for this dataset that is different from the article and also different from each distinct file. For the time being, a citation can be created with the journal article information and a link to figshare as follows:

Roggeveen, Suzanne; van Os, Jim; Lousberg, Richel (2016): Does the Brain Detect 3G Mobile Phone Radiation Peaks? An Explorative In-Depth Analysis of an Experimental Study. PLOS ONE. Dataset. https://figshare.com/articles/dataset/_Does_the_Brain_Detect_3G_Mobile_Phone_Radiation_Peaks_An_Explorative_In_Depth_Analysis_of_an_Experimental_Study_/1410034?file=6752208

Although the original citation redirects to the journal article. This article does include an embedded file that shows the data using figshare as well as DOI's for each data file separately.

2. Long-Term Preservation

As stated in the data and metadata profile, the data files use three different formats: .docx, .sav, and .sps. The .docx file format was created and is maintained by Microsoft. Files in the .docx format can be opened with word processing software such as Microsoft Word or Open Office. Although it is unlikely that this format will become completely obsolete, there is a possibility that Microsoft will transition into a different format as they did from .doc to .docx in 2007.

The .sav and .sps file formats were created to store large data and also syntax frequently to be used with IBM's SPSS statistical software. Due to this, SPSS is the most-commonly-used software to open files in either of these formats. At this moment, there is little risk that these file formats will become obsolete as it is likely that IBM will continue to use them in the present and future versions of SPSS.

An important consideration for the preservation of these data is that if they are submitted for inclusion in OpenNeuro, they would need to be converted into .json and .tsv formats. These are more widely-used formats that are even less likely to become obsolete. Considering this, it would be beneficial, for the purposes of preservation, to seek BIDS compliance for this dataset.

3. Copyright Licensing

These data are already available for reproduction for any commercial or non-commercial purposes. The current license is Attribution 4.0 International (CC BY 4.0). A better assessment of the appropriateness of this licensing would be possible if more information was available about the source of funding and their own licensing requirements. Considering the available information, the current licensing is appropriate. A different consideration would be the licensing requirements for submission to OpenNeuro. OpenNeuro requires CC0 licensing; they require that the dataset is part of the public domain. Any submission to OpenNeuro would require a change in licensing to comply with this requirement.

4. Ethical/Human Subject Considerations

The journal article related to these data states the following: "Approval was obtained from the medical ethics committee of the Academic Hospital Maastricht, on June, 6th, 2013." This would imply that the study in itself, and the data generated from the conduct of this research, complies with any applicable regulations for the protection of human subjects. The data do not include any images or any personally identifiable information from the participants of the study. Participants seem to have been assigned anonymous identifiers to protect their identity. At this moment, no changes need to be made to the data files to protect any human subjects in the underlying study.