This working paper reviews technical requirements for archiving EEG data at ICPSR including summarizing requirements for ingest, curation, and workflow management tools.

ICPSR Working Paper 3:

Technical Requirements for the DevEEG Archive at ICPSR

Amy Pienta, Ph.D, Research Scientist, ICPSR

Erin Meyer, BA, Data Project Manager, ICPSR

Kilsang Kim, MA, Business Analyst, ICPSR

Acknowledgements:

Other contributors from ICPSR to this paper include: Matt Nizol, Tamara Qawasmeh, Matthew Richardson, and Bing She. We also thank Dr. Elif Isbell, Department of Psychology, University of California Merced (UC Merced) for her leadership and contributions to the project. Funding for this pilot project was provided by the Michigan Institute for Data Science (MIDAS) Propelling Original Data Science (PODS) pilot grant program at the University of Michigan and the Inter-university Consortium for Political and Social Research (ICPSR).

ICPSR Working Paper 3: Technical Requirements for the DevEEG Archive at ICPSR

Introduction

For robust and reproducible research at the intersection of developmental science and neuroscience, there is a critical need for a dedicated data repository for developmental EEG data (DevEEG). Developmental EEG studies typically have very small sample sizes, and even though EEG is extensively used in basic, clinical, and educational research, EEG data are rarely FAIR: Findable, Accessible, Interoperable, and Reusable. The few existing neuroscience data repositories do not include standardized data formats and analysis tools, lack the proper protections and access regulations regarding the confidentiality and privacy of sensitive developmental data, and/or do not have the capability to archive and curate supplemental biological, demographic, environmental, and behavioral data that are critical to answer key developmental research questions.

DevEEG will be housed at the Inter-university Consortium for Political and Social Research (ICPSR), which provides more than 55 years of expertise in best procedures for protecting confidentiality and distributing restricted data, expertise in creating and curating metadata, and providing long term and broad access to the research community. DevEEG will be among the first data repositories in human neuroscience that can also securely archive and curate associated biological, demographic, environmental, and behavioral data. It will be the first initiative to apply ICPSR's established infrastructure for data standardization and secure data dissemination to neuroscientific data. DevEEG will also be the first EEG data repository hosted by ICPSR, expanding its data archiving and curation capabilities to neuroscience data.

To ensure widespread participation from the global developmental science community, and include users with little computer programming experience, this paper describes the technical requirements underlying implementation of the DevEEG data archive to enable easy deposit and extraction of raw EEG data and metadata.

Ingest of BIDS EEG Data

In order for any data to be archived at ICPSR, the files need to be transferred to ICPSR's systems, and metadata needs to describe the contents of and context around the data files. The standard process for making this transfer involves the depositor creating and submitting a data deposit form on the ICPSR website. We identified several requirements for the ingest process to facilitate archiving developmental EEG data. These requirements include:

1. The current data model should support the BIDS formatting standard for EEG data.
2. The deposit system must capture and make use of the full-range of study-level and BIDS-specific metadata, such as file-level metadata and organization.
3. The deposit system must be able to ingest large EEG files.

DevEEG will allow researchers to store and curate their EEG data along with any associated biological, demographic, environmental, and behavioral information collected from developmental populations. The Brain Imaging Data Structure (BIDS) is a research community standard for organizing and sharing human brain neuroimaging data (Pernet, Appelhoff, Gorgolewski, et al., 2019). BIDS provides uniform naming conventions and file hierarchies that allow for secondary users to better understand the data. In support of existing reproducible neuroscience efforts, we will use the EEG extension of BIDS (EEG-BIDS) for archiving, and during the ingest process, depositors will supply their BIDS EEG data in the form of binary and text files uploaded in zip bundles. However, the deposit system requires enhancements that will better support depositors sharing data organized into BIDS.

To add support for BIDS, we recommend adding a BIDS Importer module to the existing ICPSR Deposit Manager. This BIDS Importer will integrate tools that have been made available by the BIDS developers, such as the BIDS-Validator. The BIDS Importer would run the BIDS-Validator on the uploaded package to ensure that the data deposited are in adherence with the BIDS requirements. If the BIDS specifications have not been met (validation fails), this

module would alert the depositor and allow them to edit and update their files in the deposit system workspace. These validation results can then be passed down the pipeline to the deposit workspace and thus populate the ICPSR Metadata Editor for use by ICPSR data curators. The deposit system workspace would also still display the standard deposit module, which would allow the depositor to upload any accompanying data files that can be linked to the data in the BIDS package, if desired.

The current ICPSR deposit form includes metadata fields that are most applicable to social science survey data collection, but BIDS requires researchers to provide metadata that is not represented in ICPSR's deposit form. This metadata is not only at the study level, but also at the participant and even file level. To capture and make use of the full range of BIDS metadata, ICPSR would expand the metadata fields presented in the deposit system workspace. For optimal user experience, the deposit system could extract the computer-readable metadata from the BIDS package automatically so the depositor would not need to provide redundant information. We would likely build three schemas that define the top-level metadata for the following three object types to accomplish this requirement:

1. *EEG Data Container* describes the BIDS EEG metadata fields.
2. *Non-EEG Data Container* describes any accompanying data that is more similar to a current study (i.e., tabular).
3. *Project Container* connects the EEG and Non-EEG Data Containers so they can be presented together.

Finally, the ICPSR deposit system is not currently optimized to ingest large files, and the size of BIDS EEG packages may surpass the ICPSR import size limits (currently, 30 GB). The study that was archived during the DevEEG pilot project is an example of a small scale EEG study (n=24 study participants), and during deposit, it required a one-time increase in the deposit storage capacity before it could be submitted. It is reasonable to expect that most EEG studies will need to be able to transfer their files to ICPSR using a system that can manage

large scale file transfer. The Infrastructure Notes section describes this problem and potential solutions in more detail. The DevEEG archive will require that one or more of the proposed solutions is available to researchers for the archive to be successful.

Curation of BIDS EEG Data

ICPSR curates the data it provides to ensure compliance with the FAIR data principles. The curation process includes reviewing the data for disclosure risk, enhancing existing documentation and performing a variety of data cleaning tasks, including checking for undocumented codes and standardizing missing data. Curation staff at ICPSR will follow the detailed processing guidelines and procedures it has developed for survey data to curate any supplemental non-BIDS data, however, there are unique tasks that curation staff need to perform to ensure a deposited BIDS EEG package is suitable for archive. BIDS centers around clear and meaningful requirements for file naming conventions, file formats, and folder structure, and because of this, a deposited BIDS EEG package in general requires less curation effort. Nonetheless, we identified several technical requirements to enable BIDS EEG data to successfully move through the curation process. These requirements include:

1. Curators need access to open source tools from the BIDS ecosystem within the curation workspace
2. The metadata editor must display BIDS EEG-specific fields
3. Optional, we may want to allow curators to be able to create processed EEG data from the raw data that has been deposited.

Curators need to access several resources that were not available in the standard environment before the DevEEG pilot began. Most importantly, curators need to use the BIDS Validator to examine the deposited data and metadata. Additionally, curatorial staff need to be able to run test analyses on the deposited EEG files to make sure they were not corrupted at any point during the ingest or curation process, and this requires access to programming

languages that are able to explore EEG data, such as MatLab and/or Python. Curators who wish to examine the data using Python will need access to the MNE-Python libraries, including MNE-BIDS, while curators who prefer MatLab will need one or both of the following add-ons: 1. EEGLAB with bids-matlab-tools plugin and bids-validator extension 2. FieldTrip. These tools were not standard offerings in the secure enclave where curation takes place, and they will need to be maintained and supported by technical staff.

Since depositors of EEG data will be able to extract and supply metadata fields that are specific to developmental EEG researchers, these fields will also need to be displayed in the Metadata Editor and associated with the deposited EEG zip bundle. The BIDS metadata fields need to be built into the Metadata Editor, and if there are associated non-BIDS data, the Metadata Editor should also show those relevant fields.

Finally, ICPSR will consider developing the capacity for curators to create processed EEG data from raw data. The complex waves of raw EEG data are not interpreted easily-certainly and methods have been developed to compress, simplify, and display various “processed” summaries of the EEG. ICPSR would need to seek guidance from the scientific community on this process and develop processing standards that would be useful to advance use of data by the broad scientific community. While not essential, this step has the potential to accelerate use and impact of the data.

Dissemination of BIDS EEG Data

There is a critical need for a dedicated developmental EEG repository that can securely archive and curate EEG data along with associated biological, demographic, environmental, and behavioral information, while also ensuring confidentiality and privacy requirements of sensitive developmental data. An international user needs survey we conducted revealed that more than 50 major research laboratories from 4 continents were interested in sharing their developmental EEG data but were concerned about the confidentiality and privacy requirements. These studies

would benefit from being discoverable by secondary researchers and then available to be accessed by approved researchers. ICPSR specializes in applying the appropriate protections on data, but there are some requirements for data dissemination that the DevEEG archive should address:

1. ICPSR's search capability needs to take advantage of the range of available BIDS metadata
2. Project homepages for EEG studies will reflect additional metadata and the BIDS EEG packages
3. The ICPSR Virtual Data Enclave will be optimized for restricted EEG data access and analysis.

Researchers seeking to use developmental EEG data from ICPSR will want to be able to narrow down their search results using filters that are relevant for their area of research. To accomplish this, ICPSR's current search capabilities should expand to take advantage of the additional metadata that describes characteristics of the EEG data. This includes adding BIDS-specific metadata fields to the main search as well as building BIDS-specific search facets.

When they do not contain disclosive information, the full EEG binary files will be available for secondary users to download as a zip bundle. It would benefit researchers if ICPSR built the ability for users to choose whether they want to download the entire project (BIDS+Accompanying data) or only one of the components. Adding this flexibility will reduce any unnecessary downloads of large files and give the researcher only the files they intend to use.

ICPSR hosts a Virtual Data Enclave (VDE) that provides access to restricted-use data in a secure environment. The VDE is a virtual machine launched from the researcher's desktop that operates on a remote server. The virtual machine restricts users from printing, emailing, copying, or otherwise moving files outside of the secure environment, either accidentally or intentionally. Software tools and support available for use within the VDE include geospatial

analysis tools, statistical analysis software, and various documentation programs. In its existing state, the VDE has not been optimized for use by developmental researchers. ICPSR should consider evaluating data user requirements to design an optimal VDE experience for accessing and analyzing EEG data. It is likely that programming languages like Python and MatLab will require additional add-ons similar to the ones put in place for Curators. The computational power and storage available in the VDE are also anticipated limitations for researchers.

Workflow Management of BIDS EEG Data

The three aforementioned stages, ingest, curation, and dissemination each have individual requirements for how to add support for the DevEEG archive at ICPSR, but each of these systems are also connected. As ICPSR continues to evolve its data model to accommodate different types of data and archival projects, we will need to add support for the BIDS EEG data object type and define a supporting workflow through our systems. To accomplish this, we propose building three services that will enable the workflow management of BIDS data:

1. EEG-API
2. EEG-Manage
3. EEG-Web

The EEG-API provides a protocol for communication between the various ICPSR apps (including Curation Manager) that make up the repository system. The EEG-API would know that the object being passed around is a BIDS EEG data and know the rules around how we manage this object.

The EEG-Manager defines a set of rules that define what will happen to the BIDS EEG object type. It will offer the broader workflow and functionality that will plug the BIDS EEG objects into Curation Manager. The EEG-Manager will need to be triggered either by information supplied by the depositor or by ICPSR staff that an object is indeed a BIDS EEG object type.

After a BIDS EEG deposit has gone through the proper steps and is ready to be disseminated, the EEG-Web service will display the final product via a project homepage. This display should be able to render differently for different types of deposits, such as a BIDS-only project homepage can look different than a BIDS+Accompanying Non-BIDS Data project homepage.

Challenges

Several problems that were highlighted when we tested archiving EEG data at ICPSR were not due to unique characteristics of EEG data or of BIDS. Instead, the testing put a spotlight on general infrastructure limitations that ICPSR should prioritize tackling as it grows to accept more big data and new data types. The following challenges describe infrastructure improvements that will address persistent issues that are likely to recur for other exploratory projects.

As described earlier, EEG deposits will very often exceed the standard size limits for data deposits. As ICPSR runs into use cases where large files need to be securely transferred into our custody, there is a growing need to make the ingest process more powerful and flexible. Adding support for multiple file transfer methods will aid the ingest process and allow ICPSR to take on more projects that involve big data files. The standard ingest process using the deposit system web GUI should be made more robust to support the upload of larger files (> 30GB). These improvements should also offer a way for upload interruptions to be communicated to the depositor and recovered. While improving the web GUI would help ease some pain points associated with uploading files to the deposit system, we should also explore solutions that use clients like SFTP and/or Rclone to transfer large files securely into our environment. This would enable the transfer of much larger data packages into our systems and could also leverage the Cloud.

Additionally, some of the roadblocks are related to ICPSR using Fedora for the storage solution during the ingest step. ICPSR should consider transitioning off of Fedora and using S3, which has several advantages. If S3 were in place, data files that were transferred into ICPSR's systems using one of the above ingest methods (web GUI, SFTP, Rclone) would be moved into a File Drop space on S3, which is what ICPSR currently uses for Cloud storage space. This intermediate location then would be able to house the data until the files are moved into the ingest pipeline. The S3 storage solution also could manage versioning and archive, which would further reduce and eventually eliminate ICPSR's reliance on Fedora.

Ingest is not the only point of the pipeline that currency struggles to handle large files. The dissemination stage also depends on the speed of the internet speed of the user downloading data. ICPSR should consider implementing functionality where users can download data from outside the web browser. Existing technologies like Node.js, S3, or DataLad are worth exploring, and if implemented, these solutions could be recommended to users who are requesting downloads that are at higher risk of interruption. A key benefit of these solutions is they all offer ways for download interruptions to be recovered. Alternatively, ICPSR could experiment with breaking up zip bundles into reasonable download sizes to support downloading large files from within the web browser.

Conclusion

The DevEEG archive can address many of the needs of the developmental research community by leveraging the long-standing infrastructure at ICPSR, but integrating the BIDS EEG data type into ICPSR's current technology platform also comes with unique and challenging needs. This paper explored these challenges and suggested possible solutions, and when addressed, the DevEEG archive will serve researchers seeking a secure data repository for developmental EEG data.

Bibliography

Pernet, C.R., Appelhoff, S., Gorgolewski, K.J. *et al.* EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci Data* 6, 103 (2019). https://doi.org/10.1038/s41597-019-0104-8