Data and Services Discovery projects - Transformative Data Collections

Standardisation of protocols for collecting linked images and electrophysiological data from in vitro and in vivo studies on neural cells and tissues

Approach

Activities.

1. The following questions were addressed.
    a. Understand how neuroscientists and neural engineers collect, store, analyse and archive in vitro and in vivo neural data.
    b. What are the key open databases for sharing images and electrophysiological information?
    c. How do users interact with these databases, including any standardisation protocols?
    d. What are key needs for improving standardisation of neural data collection?
    e. How can we support more effective accessibility and sharing of neural data?
    f. What types of platforms could we use to support “data interoperability”
    g. What mechanisms will be put in place for secondary use of research data to ensure privacy and security of data
2. A review of the open platforms available was conducted, and a summary of existing databases and search tools in the neural imaging and electrophysiology space was completed.
3. A review addressing the key questions above was also conducted via interview of key neuroscience and neural engineering groups nationally. Ethics approval for the interviews was obtained on 23/07/2019 via the UNSW Human Ethics Panel (HREA approval number HC190549) which delayed the start of the project by approximately 7 weeks.
4. Two research assistants, one in Sydney and one in Melbourne were recruited to conduct interviews and a total of 19 interviews were completed.
5. A workshop was held in Sydney on February 13, 2020 to discuss all findings and to specifically focus on questions e. to g. above. Participants included 3 researchers from Melbourne and 4 researchers and 1 IT infrastructure manager from Sydney.
6. The final report (submitted February 21, 2020) was completed based on the compiled findings from the surveys, web platform review, and the workshop.

Participants and collaborators

Name	Institution	Role
Laura Poole-Warren	UNSW Sydney	Lead CI
David Tsai	UNSW Sydney	Deputy Lead
Ulises Aregueta	UNSW Sydney	Researcher
Nigel Lovell	UNSW Sydney	CI
Grant Kelly	UNSW Sydney	RA* - Sydney based
Luc Betbeder	UNSW Sydney	RI** support
Penny Martens	UNSW Sydney	CI
Mohit Shivdasani	UNSW Sydney	CI
David Grayden	UoM	CI
Nina Erfanian	UNSW Sydney	RA – Melbourne based

Mirella Dottori	UoW	CI
Sally McArthur	Swinburne	CI

*Research Assistant; **Research Infrastructure

Participants were recruited under the informed consent conditions of the UNSW Ethics approval (HC190549). Their names are not included in this report; however, researchers from the institutions in the Table below were interviewed.

Institution	**Position**	**Individuals**
UNSW	Research Associate	1
	PhD Candidate	2
	Academic Staff	5
Western Sydney University	Academic Staff	1
University of Sydney	Academic Staff	1
The Bionics Institute	Researcher	3
	Professional Staff	1
Florey Institute	Researcher	2
NVRI	Researcher	2
University of Melbourne	Academic Staff	1

Outputs

To date there have been no outputs produced.

FAIR

The project has contributed to understanding the challenges presented in making the types of data being studied more FAIR. DOIs exist for some public databases (eg: MGH/MF Waveform Database - https://doi.org/10.13026/C26K5Q) and are described by metadata records, the identifiers are not included in all metadata records. These are mostly public data sets used for collaboration by researchers. However, much of the imaging and electrophysiology data in the neuro area are still captured locally and have not been published.

The FAIR assessment spreadsheet attached summarises findings from this project. The FAIR framework will be leveraged to inform the development roadmap for these datasets.

Collaboration and coverage

Through our survey interviews, we achieved broad coverage of relevant researchers and professional support staff in NSW and Victoria. Interviews conducted indicated that equipment used is highly variable and dependent on the researcher's preference and current de facto standards used in the majority of laboratories world-wide. The key interactions with global datasets was through the review conducted of the open access databases in the field. A symposium at the Shine Dome in Canberra on October 9, 2019 *Data sharing: neuroscience, microscopy and experiments symposium,* was hosted by a related ARDC grant (Poole-Warren, Tsai and Kelly attended). At this workshop there was an opportunity to engage with international leaders of relevant open data sharing platforms. This informed our ongoing review of these platforms, which was discussed at our local workshop in February 2020.

Sustainability

The project team closely working with other ARDC Discovery teams to share results and work towards synergistic, sustainable outcomes. As noted above, three team members attended the October 9 symposium which supported understanding of how to enable interoperability and collaboration between neuroscience and neurotechnology researchers.

Project Lead CI Poole-Warren is a member of the successful Australian Brain Alliance (ABA) ARDC Discovery Project focusing on establish and operate an Australian Brain Data Commons (ABDC). With the support of the Academy of Science, ABA is developing a three-year plan, including funding proposals that will ultimately provide sustainable infrastructure to support the range of collections in Brain science and technology. Our ARDC Project research on standardisation will provide input to the development of the ABDC which will play a key role in ongoing work in this space.

Learnings

Key open databases for sharing images and electrophysiological information

Background research was performed on current neuroscience databases with a focus on electrophysiology and imaging data. In accordance to the FAIR principles a search was conducted to identify databases that are Findable, Accessible, Interoperable and Reusable. Online platforms such as the Neuroscience Information Framework (NIF), Mendeley data and the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC), provide with search engine tools to access neuroscience databases.

Over 270 databases were identified that share neuroscience data. However only ~0.03% and 0.16% correspond to datasets related to electrophysiology and imaging, respectively. The most frequented databases for neuroscience studies that include access to electrophysiological and neuroimaging data are listed below. These are open databases that provide access following free online registration.

- CRCNS - Collaborative Research in Computational Neuroscience
- Neuroelectro.org
- ABA Mouse Brain: CellType EphysData
- Canadian Open Neuroscience Platform(CONP)
- OpenNeuro
- Physionet.org
- Neurosynth.org
- Neuromorph.org
- Hippocampome.org
- Neurovault
- Neurodatabase.org

Data within these databases can be processed data (tables and figures) reporting mean values of cell properties or parameters or raw data. Raw data is usually presented in a Hierarchical Data Format (HDF file). This file type allows storing multidimensional arrays or datasets in groups of headings that describe experimental parameters, data properties and general information regarding data acquisition. These groups of headings within the file present an organised metadata. HDF files are commonly used by proprietary software to save data acquired from electrophysiological experiments. They can also be created by the user using programs such as MatLab, Python, R or computer programming language such as Fortran or C++. Although useful, creating and reading the file from databases requires programming skills and understanding how HDF files are organised. Also, the information included in the groups of headings and naming is not standardised.

Collaborative efforts have been created to address this issue such as the initiative of Neurodata Without Borders (NWB), which aims to provide a platform to standardise neuroscience data on an international scale. This consortium offers tools that convert datasets into standardised NWB files with specific field names that will allow for a better data description and interpretation. However, in most cases, the information available to describe the data relies on the metadata

provided by the authors sharing the dataset. Thus, there is a need to develop new data naming standards and strategies to encourage researchers to apply them when sharing their data.
Although most databases are accessible and relatively easy to find, processing data is limited by software accessibility and the user's skills to operate such software. Also, reusability of data appears to be of concern since, while most data is linked to a peer-reviewed article, this does not guarantee the validity of the data. These limitations further complicate the interoperability. There is a broad range of tools and platforms that have been developed to support data sharing and access. Similarly, a gallery of software allows creating and reading data from complex data format types such as HDF. However, this requires programming knowledge and skills. Also, to enable trustworthy data sharing new standards for data file naming and classification are required.

Summary of Survey Findings

Format: Data format tend to be proprietary for the equipment used or a de facto industry standard (eg: TIF for images). Databases are typically not used to store raw data. Generally, simple file/folder structures but with elaborate filenames to indicate experiment and key parameters. Metadata and experimental protocol are generally stored in a lab book or electronic file (Word, Excel). Some experimentation with electronic lab books.

Linking all the raw data and metadata that is located in different files and folders to generate a 'coherent story' of the experiment is noted to be an issue. As a result of the above the ability to replicate and experiment becomes problematic to anyone other than the person(s) familiar with the protocol. Data integrity and archiving is of key importance. Generally, this is done using a lab PC, and/or a server with IT support. Often separate HDD copies are kept. Increasingly large volumes of data (up to 100's of GB per experiment) are becoming an issue. No use of cloud storage due to this.

Analysis: Data analysis tends to be done using standard software including MATLAB, Julia, Labchart, Prism, R, and Excel. Once again databases are not typically used. Instead, file systems are the norm. Results are often tabulated in Excel and working copies of raw data may be manipulated by multiple researchers making the file system complicated to maintain. Generation of well documented scripts (eg: MATLAB) can aid in understanding and replicating analysis. An interesting case was the use of 'R Studio' to provide a document that included raw data ref, analysis scripts, results, comments and discussion to provide a complete review of the statistical analysis protocol.

Key Lessons Learned

1. Given that imaging and electrophysiology data is obtained in a wide range of formats, and storage tends to be in local, private storage or institutional archives, the current potential for wider national sharing is limited.
2. It is not practical to standardise all the different data formats, however a standard method of storing and relating these data pieces (metadata) could aid data interoperability, sharing, and experimental understanding and replication.
3. It is important that standard analytical tools accept all relevant raw data formats so that the analysis outputs are similar, and the tool selected remains the researchers' choice.
4. It is desirable for a third party to be able to follow the steps of the analysis (ie: the 'analysis story') without having to review all the code. An overall protocol document (metadata) would again aid data interoperability, sharing, and replication.

ARDC could support a national approach to achieving openly available metadata through working with groups responsible for research infrastructure and institutional repositories. This approach could allow local/private storage but institute a secondary use of deidentified data structures (for example E-Research Institutional Cloud Architecture. (*ERICA*) – Professor Louisa Jorm). Such cloud-based infrastructure could recommend or require common storage formats.

International approaches to unifying the availability of metadata could be achieved through working with publishing organisations and international professional societies and other bodies.

Impact

Research Outcomes

The outcomes of this project will contribute nationally through input to the ABA ARDC project as outlined above. Our group have expertise in design of database systems that comply to regulatory requirements (FDA and TGA) and meet appropriate privacy and security frameworks. Secondary materials include the design architecture of a secure public cloud-based storage system (with potential to be HIPAA compliant) that leverages data from multiple sources and facilitates data access and interoperability using standard health messaging (FHIR).

Broader Impact

In a workshop at the recent IEEE EMBS Neural Engineering conference (NER'19) in San Francisco (see https://neuro.embs.org/2019/workshops/from-flexible-materials-to-cell-scalerecording-emerging-frontiers-in-neural-interface-technology/), Thomas Steiglitz conducted a panel discussion on community needs in the cell scale recording space. This was attended by over 100 experts in neural engineering from across the globe. CI Poole-Warren was a panel member at the workshop. One of the key discussion points was the need for data standardisation to underpin sharing and collaboration by ensuring that data was collected using consistent protocols and was searchable and accessible to the community. This would benefit the research community and industry supplying relevant imaging and electrophysiology equipment.

Report prepared by: Laura Poole-Warren and Ulises Arequeta, UNSW Sydney
Date: 21/02/2020