Discovery Informatics: AI Takes a Science-Centered View on Big Data
AAAI Technical Report FS-13-01

Creating an Urban Legend:
A System for Electrophysiology Data Management and Exploration

Anita de Waard[1*], Jeremy Alder[2], Shawn D. Burton[3,4], Richard C. Gerkin[3,4], Mark Harviston[1], David Marques[1], Shreejoy J. Tripathy[4,5], Nathaniel N. Urban[3,4]

[1]Research Data Services, Elsevier, US, [2]User-Centered Design, Elsevier, UK, [3]Department of Biological Sciences, [4]Center for the Neural Basis of Cognition, and [5]Program in Neural Computation, Carnegie Mellon University, Pittsburgh, PA;
*Correspondence: a.dewaard@elsevier.com

Abstract

We have created a tool to identify and store experimental metadata during the execution of an electrophysiological experiment, and a semantic architecture to enable access, manipulation and integration of this data to support a collaborative research environment. We discuss possible extensions of this work to aid data sharing and semantic research frameworks.

Motivation

Understanding the electrical activity of the human brain is a leading focus in today's scientific landscape (Alivisatos et al., 2012; Abbott, 2013; Insel et al., 2013; Pastrana, 2013). Electrophysiological experiments in model systems, such as the rodent brain, are central to these research efforts. At its core, neuronal electrophysiology involves the study of ion channels and the emergent electrical properties that channels imbue neurons and neuronal networks with (Hille, 2001). Understanding each ion channel and electrical property requires years of study. Thus, translating specific data about individual channels, neurons, and networks into knowledge about brain-wide activity and function requires effective sharing and integration of data across a great number of experiments and laboratories. A system capable of complete and effective integration of such data is not currently available.

One of the greatest impediments to effective integration of electrophysiological data is that ion channels, and consequently neurons and neuronal networks, are tremendously sensitive to a host of experimental factors. Such factors – the experimental *metadata* – include (but are not limited to) model species (*e.g.*, rat vs. mouse), strain (*e.g.*, Sprague-Dawley vs. Wistar), and age (*e.g.*, perinatal vs. adult); experimental temperature (*e.g.*, room vs. physiological temperature); solution composition (*e.g.*, high vs. low intracellular chloride concentrations); and experimental equipment (*e.g.*, sharp vs. patch electrodes), and vary both within a laboratory – from experiment to experiment – and between laboratories. Even if two laboratories are willing to share their data, lack of sufficient and standardized metadata attached to each datum precludes accurate interpretation and effective integration. Currently, this metadata is most often logged in laboratory notebooks during or after each electrophysiological experiment, and is therefore not easily compiled or shared. Thus, to apply a new analysis to published data, it is currently easier to repeat the published experiment than to attempt to acquire the published data and associated metadata. This practice is clearly insufficient if we are to achieve the large-scale, effective integration of electrophysiological data needed to map and understand human brain activity (Insel et al., 2013).

To overcome these limitations, we here describe a system that enables the collection of metadata in a digital, standardized format by the experimentalist, during the course of an electrophysiological experiment. Current efforts in semantically enabled Electronic Lab Notebooks abound, (see e.g. in Gil et al., 2011, Freire, et al., 2012; Frey, 2009; Talbott et al., 2005, and many others). It is not our goal, nor is it within our ability, to add yet another entry to this already well-stocked field. But large-scale adoption of these - often excellent – tools by bench researchers, in particular in biology and the neurosciences, has so far been lacking. In creating a rather simple application, but tailoring it completely to the needs of a specific lab we can remove some possible barriers to adoption, by being very close to the user's workflow; if nothing else, we hope that our project can help elucidate the reasons behind the reluctance in using workflow tools.

Our system, dubbed Urban Legend (after the Urban Lab[1] with which it was co-developed), is an electronic

[1] https://www.bio.cmu.edu/labs/urban

laboratory notebook application that runs on (tablet) computers and smartphones, and integrates with established electrophysiology acquisition software, Igor Pro[2] to automatically synchronize experimental metadata with electrophysiological data on a metadata server. In this paper, we describe the system we have built to date, and our plans for future developments.

Our System

Our system consists of five key components (see Fig. 1):

- Data Entry App
- Metadata Database
- Igor Pro Integrator
- Ontology Integration
- Data Dashboard.

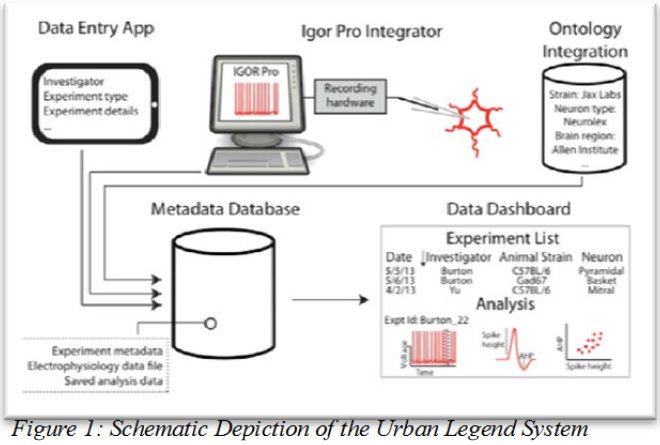

Figure 1: Schematic Depiction of the Urban Legend System Components.

These will now be discussed in turn.

Data Entry App

In building the App to capture the experimental metadata, we had three goals:

1. Make it easier to capture metadata digitally than with any other means (paper lab notebook, web forms), thereby reducing barriers to early and consistent electronic metadata entry;
2. Give value back to the research lab by moving to a digital-first experimental description and thereby facilitating compilation and sharing of research data;
3. Reduce the effort to create a highly lab-customized metadata capture system to help more rapid adoption.

To achieve these goals, we built a data entry App that runs on any tablet device, and provides multiple mechanisms to create lists and generate automatic layouts for data entry. The App can be customized either through a single data descriptor text file or through the metadata server, and changes to the App are applied instantly and automatically through the server. Because the App is continuously synchronized with the metadata server, the descriptions of all experiments are available to anyone in the laboratory as soon as the experiment is done, providing immediate value to the metadata captured.

Metadata Database

The metadata database facilitates interoperability between all components: the electrophysiology acquisition software (written in Igor Pro), the Data Entry App, and the Data Dashboard. The only time a person directly interacts with the server is when they configure drop-down choices for the App or add new investigators to the list. The server provides the investigator list to the App and to Igor Pro, and the active run information to Igor Pro. The metadata server, which is hosted in the cloud, saves all experimental data and metadata from both Igor and the App so it can be accessed by the Data Dashboard.

The metadata server consists of 2 pieces: a PostgreSQL[3] database and a Django[4] web-application. The PostgreSQL schema is very simple: each metadata entity is identified by a type-4 (randomly generated) UUID[5] and a scope string, and stores the investigator, created and last modified timestamps, and a set of key-value pairs stored using the PostgreSQL hstore[6] extension. Investigators are stored in a second table with an ORCID[7] ID and a display name (ORCID IDs are used to uniquely represent investigators throughout). A third table stores the choices for drop-downs in the App. The Django web-application is an HTTP API with simple API key based authentication and JSON[8] as the transport format. API endpoints (also knows as Network API Calls) currently exist to:

- get a list of investigator's ORCID IDs and display names
- get the drop-down choices for a specific investigator
- get experimental metadata,
- save experimental metadata
- activate a run, and
- get the active run.

Igor Pro and the App can use these endpoints to coordinate efforts (the App activates a run, and Igor Pro gets the active run), as well as – through the metadata server - store metadata to the cloud. Since metadata is stored as a simple key-value pair format it is up to those programs to keep the data consistent and clean.

[2] http://www.wavemetrics.com/products/igorpro/igorpro.htm

[3] http://www.postgresql.org/
[4] https://www.djangoproject.com/
[5] http://tools.ietf.org/html/rfc4122
[6] http://www.postgresql.org/docs/9.2/static/hstore.html
[7] http://orcid.org
[8] http://tools.ietf.org/html/rfc4627

Igor Pro Integrator

Although in principle Urban Legend could leverage any software interface between the experimentalist and her preparation, we demonstrate the project using Igor Pro[9], a program with strong community adoption in neuroscience, as well as in physics, chemistry, geology, and meteorology. It consists of a programmable GUI, a statically-typed scripting language with a compiler and debugger, and low-level extensibility via modules written in C/C++. The latter feature has been used previously to enable Igor Pro to control and acquire data from a wide range of instruments, including digitizers, cameras, and other devices used in electrophysiology. The former features make control of and acquisition with these devices transparent and flexible, as well as enabling data analysis and the production of publication-quality graphics.

Within Igor Pro, there is a choice of software packages implementing electrophysiology-specific requirements including communication with specific instruments and graphical presentation/control of outputs/inputs meeting the needs of the experimentalist. To enable extraction of data and metadata from saved experiment files, any such package needs to implement an interface to the App and the metadata server.

We define this interface generally, and implement it for two choices of Igor package:

(1) "Recording Artist[10]", a package developed by one of us [RG] and used in over a dozen laboratories worldwide;

(2) Nathan Urban's implementation of electrophysiology data acquisition in Igor Pro, a custom package developed by Nathan Urban for use in his laboratory.

For the Urban Legend project, use of Igor Pro is unchanged for the experimentalist, with the exception that a profile must be selected when Igor Pro is first opened to obtain the Urban Legend profile settings and coordinate information about experiment status. Everything else happens "behind the scenes", including the following stages:

(1) Igor Pro periodically obtains information about the status of the current experiment, as indicated in the App interface. This status is summarized by the experiment scope in the App.
(2) The scope is bound to Igor Pro data objects, so the scope under which each object was created (*e.g.*, each "sweep" was collected) can be determined later.
(3) Optional experimental metadata can be posted by Igor Pro to the app.
(4) The completed experiment is exported in HDF5 format and uploaded to the server.

Once the HDF5 file is on the server, experimental metadata is programmatically extracted, using the interface described above. We chose to use HDF5[11] to store electrophysiology data files in part because of its recommended use by electrophysiology data standards group[12].

Since different Igor Pro packages may store acquired data in different locations within the hierarchical Igor Pro file structure (which is inherited precisely by the HDF5 file), the interface must specify where particular pieces of data, and associated metadata, will be found within the file. For example, one interface method is GetSweeps(scope), which returns an array of HDF5 file locations corresponding to sweeps collected under a given scope. Some of the extracted metadata is immediately stored on the server in a relational database, while other metadata and all data remains in the HDF5 file to be requested on demand. With this architecture, an investigator wishing to visualize data matching certain metadata attributes can do so quickly by using the Data Dashboard to constrain the search and explore it.

Ontology (Semantic) Integration

An essential part of the Urban Legend project is ensuring that the collected data remains clear and understandable throughout the research life cycle. This is helpful to scientists when they are reviewing their own data as well as to other scientists or collaborators who wish to view or reuse this data.

To this end, we have taken the preliminary step of semantically marking up the entities collected via the metadata app with unique resource identifiers (URIs) and externally referenced definitions (a preliminary version is available here[13] - future versions will be made available through [14] and [15]). For example, to refer to specific strains of genetically modified mice, we use URIs provided by Mouse Genome Informatics (MGI). Similarly, when referring to electrophysiology specific concepts, such as electrode type or recording configuration, we intend to use identifiers provided by the Ontology for Experimental Neurophysiology (OEN[16]). We can further use the OEN to annotate the recorded electrophysiological data with neuron-specific measurements such as action potential amplitude and width.

In searching for unique identifiers for each of the concepts that we were representing through our metadata app, we often encountered concepts with no suitable external identifier when using OntoBee (Xiang et al, 2011) to perform term searches across ontologies. This inability

[9] http://www.wavemetrics.com/products/igorpro/igorpro.htm
[10] http://bitbucket.org/rgerkin/recording-artist
[11] http://www.hdfgroup.org/HDF5/
[12] http://datasharing.incf.org/ep/HDF5_data_standard
[13] https://docs.google.com/spreadsheet/ccc?key=0AoqX2haLXhtLdHdlZ0ZqaGtOQS0yRUppNWdFMnZ2eHc#gid=0
[14] http://www.neuinfo.org/vocabularies/
[15] https://code.google.com/p/eagle-i/
[16] https://github.com/G-Node/OEN

to find terms in an existing ontology was particularly troublesome for neuroscience concepts like specific neuron types (e.g., olfactory bulb mitral cell) or methodological details (e.g., sagittal brain slice). Thus we are working with ontologists at eagle-i[17] (M. Haendel and N. Vasilevsky) to add these terms to a suitable ontology (e.g., olfactory bulb mitral cells would be added to the Cell Ontology[18]).

We acknowledge that simply annotating concepts and storing URIs into our PostgresSQL database is not equivalent to representing collected data using a formal ontology. We are implicitly encoding relationships and sub/superclasses through the structure and scoping of our metadata app and database, as opposed to explicitly using an ontology. Thus we cannot currently export our data using RDF, which limits the potential power of this resource in the short term. However, with our current implementation we can expose the data to aggregators such as the Neuroscience Information Framework (Gardner et al, 2008). Thus, if a user searches for information about a specific neuron type within the NIF portal, data from our recordings could be returned. Moving forward, we plan to continue collaborating with the eagle-i ontologists to implement a formal ontology model for this data and thus expose our data using linked data principles as RDF.

Data Dashboard

The final planned outcome of this system is a tool that will help the PI, and everyone in the laboratory, to assess the results of the collected experimental results and analyze them to find correlations and connections between different results. Since all metadata will be stored in a server and normalized to a series of ontologies, it will be possible to do this work locally and remotely, and enable collaborations with other groups to allow integration with other data sources, such as NIF (Gardner et al., 2008) and Neuroelectro[19].

The Data Dashboard is currently in the planning stages, but the plan is to have it contain four different steps:

1. **Pick:** this step will allow the researcher to select the experiments of interest, using search and browse functionalities working on the metadata collected in the App and the Igor Pro Integrator;
2. **Process:** this step allows the researcher to analyze the experiments using mathematical tools (such as MatLab[20] or Mathematica[21]) to find derived wave properties of interest, and associate them with an experimental data set or collection;
3. **Plot:** this step allows the researcher to plot, compare, and analyze the processed data and see correlations, overlaps and other connections between different data sets, ordered by metadata;
4. **Publish**: this step allows export of any of the graphs/plots/data generated above into a publishable format (e.g. PPT/Word images, other image formats, or Computable Document Format[22]) while staying connected to the raw data and metadata stored in the experiments.

Next Steps

Although tailored to the unique needs and workflows of the Urban Lab, we believe the approaches, technologies used and interfaces developed for the Urban Legend project are robust and malleable enough to be applied to other experimental settings and environments. We will be making all of our software, vocabularies and metadata standards available in open source, so these can be used to build similar systems in other research environments. In particular, we believe that the conceptual system outline – the five components described above – can be a useful architecture for any system used to store, manage and analyze research data. Integration with other tools is an exciting direction that we are eager to pursue; *e.g.*, we believe there are great possibilities for integration of these components with laboratory information management systems, workflow tools, institutional data repositories and authoring and editing tools.

A lack of willingness to share data with the world at large is a known inhibitor for use of research data management systems (see e.g. Borgman, 2012). That is why an essential aspect of this system is that, although the architecture and standards used are open and interoperable with current Linked Data models, in our project the researchers themselves stay in total control of the data at all times. It is up to the individual researcher (the Principal Investigator, or head of the laboratory) to decide whether any of this data is shared outside of the laboratory. Therefore, we expect the barrier to adoption to be lower than that of fully 'open' solutions such as FigShare[23] or academic data repositories such as DataVerse[24], Dryad[25], or many others (see e.g. DataBib[26] for a listing).

We hope and expect that this system will save time and, more importantly, lead to important scientific discoveries that could not be made if the experimental data remained cloistered on individual's hard drives, or the metadata

[17] https://www.eagle-i.net/
[18] http://cellontology.org/
[19] http://neuroelectro.org/
[20] http://www.mathworks.com/products/matlab/
[21] http://www.wolfram.com/mathematica/
[22] https://www.wolfram.com/cdf/
[23] http://figshare.com/
[24] http://thedata.org/
[25] http://datadryad.org/
[26] http://databib.org/about.php

inaccessible in paper notebooks. In particular, we expect that adoption of the Urban Legend App will facilitate biological discovery in at least two specific ways.

First, automatic integration of experimental metadata with electrophysiological data will substantially streamline the collection and comparison of findings across experiments and experimenters. Immediate access to such metadata as animal age, strain, and sex will allow experimenters to more soundly match complementary data or to account for any differences. For example, complex features such as a neuron's excitability may differ between experiments due to a direct age-dependent decrease in input resistance (e.g., see: Zhu, 2000).

Second, more systematic collection of experimental metadata will allow researchers to discover otherwise unexpected relationships between experimental conditions and electrophysiological properties. For example, input resistance is expected to decrease with age due to neuronal growth and insertion of more channels (Hille, 2001), but other properties, such as resting membrane potential, can also show a less intuitive age-dependence (e.g., see: Zhu, 2000). Currently, individual studies can make such unintuitive discoveries with careful collection of experimental metadata. Generalization of such findings to multiple brain regions and many neuron types, however, is beyond the scope of any individual or small group of investigators, and will require systematic comparison of experimental metadata and electrophysiological data. In the end, we expect that these possibilities will drive researchers to open up their data to others, and demand that other's data be accessible to them.

Together with government initiatives such as the recent OSTP mandate for open access of publications and data (Holdren, 2013) it seems inevitable that the drive towards making research data open and accessible will continue to grow. The Urban Lab, for one, will be prepared.

Acknowledgements

We gratefully acknowledge the comments from our anonymous reviewers, and fruitful discussions with Melissa Haendel of Eagle-I and Anita Bandrowski of NIF.

This project is a collaboration between Elsevier Research Data Services and the Urban Lab, and is fully funded by Elsevier. All vocabularies developed during the course of the project will be made openly available through the NIF[14] and eagle-I[15] portals, and all software will be made available through the project website[27].

[27] http://researchdata.elsevier.com/urbanlegend

References

Abbott A (2013). Brain-simulation and graphene projects win billion-euro competition. Nature News, 2013: doi:10.1038/nature.2013.12291

Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R (2012). The brain activity map project and the challenge of functional connectomics. Neuron 74:970-974.

Borgman, CL (2012). The Conundrum of Sharing Research Data, Journal Of The American Society For Information Science and Technology, 63(6):1059–1078, 2012.

Freire J, Silva, CM (2012). Making Computations and Publications Reproducible with VisTrails. Computing in Science and Engineering 14(4): 18-25 (2012).

Frey, JG, The value of the Semantic Web in the laboratory, Drug Discovery Today, Volume 14, Issues 11–12, June 2009, Pp. 552-561, http://dx.doi.org/10.1016/j.drudis.2009.03.007.

Gardner D, Akil H, Ascoli GA, Bowden DM, Bug W, Donohue DE, Goldberg DH, Grafstein B, Grethe JS, Gupta A, Halavi M, Kennedy DN, Marenco L, Martone ME, Miller PL, Müller HM, Robert A, Shepherd GM, Sternberg PW, Van Essen DC, Williams RW (2008). The neuroscience information framework: a data and knowledge environment for neuroscience, Neuroinformatics. 2008 Sep;6(3):149-60. doi: 10.1007/s12021-008-9024-z. Epub 2008 Oct 23.

Gil, Y. Ratnakar, V., Kim, J., Gonzalez-Calero, P.A., Groth, P., Moody, J., and Deelman, E. (2011). 'Wings: Intelligent Workflow-based Design of Computational Experiments', *IEEE Intelligent Systems*, 26, 62–72.

Hille B (2001). Ion Channels of Excitable Membranes, 3rd Ed. Sinauer Associates: Sunderland, MA.

Holdren, JP (2013). Memorandum re. Increasing Access to the Results of Federally Funded Scientific Research, Feb. 23, 2013, http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

Insel TR, Landis SC, Collins FS (2013) The NIH BRAIN Initiative. Science. 240:687-688.

Pastrana E (2013). Focus on Mapping the Brain. Nat Methods 10:481

Talbott T et al (2005). Adapting the Electronic Laboratory Notebook for the semantic era, International Symposium on Collaborative Technologies and Systems, Proceedings (CTS 2005) (2005), pp. 136–143.

Xiang Z, Mungall C, Ruttenberg A, He Y. (2011). Ontobee: A Linked Data Server and Browser for Ontology Terms. *Proceedings of the 2nd International Conference on Biomedical Ontologies (ICBO)*, July 28-30, 2011, Buffalo, N