Ontology based data architecture to promote data sharing in electrophysiology*

Brenda Farrell
Bobby R. Alford Department of Otolaryngology – Head & Neck Surgery Baylor College of Medicine
Houston, TX USA
bfarrell@bcm.edu

Jason Bengtson
Kansas State University Libraries
Kansas State University
Manhattan, KS, USA
jbengtson@ksu.edu

Abstract— Strategies to improve the preservation, searchability, and discoverability of research data are a priority. To facilitate these efforts in cell electrophysiology and biophysics we propose that ontologies be used to design and annotate data, as they provide a substantive metadata structure, with reasoned-definitions arranged in a logical, hierarchal structure where the meaning of data are unambiguously assigned. We illustrate this by describing our cell electrophysiology data with an ontology. We then make this hierarchal structure with definitions the basis of the data architecture which is implemented upon transforming the data into the storage format: Hierarchical Data Format version 5 (HDF5).

Keywords— big data, data management, HDF5, auditory, outer hair cell, OHC, application ontology

I. Introduction

Data were inconsistently reported in the past, although data forms the back-bone of much of scientific discovery. There was no motivation for researchers to develop intuitive human-understandable (and machine readable) data structures that both the public, and their scientific peers, could readily access. The limits of this approach for many scientific disciplines have revealed a number of deficiencies, including the inability to reproduce key findings, coupled with excessive (additional) costs to the tax payer [1]. The need to provide data sets that support key findings of research is particularly relevant in fields where data collection is slow and requires the sacrifice of mammals. This is the case in auditory electrophysiology. To facilitate data preservation and sharing, we describe our effort to transform electrophysiological data from private to public use.

II. Description of Data

The data were generated by whole-cell voltage clamping isolated outer hair cells obtained from the domestic guinea pig (i.e., cavia porcellus). The electrical properties (e.g., membrane capacitance) of the outer hair cells were then determined [2] from the electrical recordings. The data encompass results from the assay, as well as the associated properties of the animals and cells; the experimental conditions employed; and a description of the devices used.

III. Use Ontology to Describe Data

To provide for the durability of the data across both time and space so that others may reasonably expect to make use of it, we describe the data with an ontology; essentially, a structure which places defined concepts in a logical relationship with one another as a way to describe things, events, or ideas that exist in the objective world. Many ontologies exist, including a large number which focus on various aspects of the biological sciences. Of relevance is Ion Channel Electrophysiology ontology, ICEPO [3] that describes concepts that are associated with electrical and temporal characteristics of voltage-gated ion channels. Although some of the concepts (e.g., gating current, ICEPO_0000049) have been used to describe the electrical characteristics of outer hair cells, the membrane protein that forms part of the voltage-sensing component in the lateral membrane of outer hair cells is not an ion-channel. We generalize, where appropriate, the concepts introduced in ICEPO to align them with voltage-dependent behavior of membrane assemblies. Another more extensive ontology is the Ontology of Physics for Biology (OPB) [4] that was developed to annotate computational models of biological systems. It uses physical (e.g., thermodynamic) dependencies to describe biological processes, and although it also covers some relevant concepts, it does not describe the design, collection and

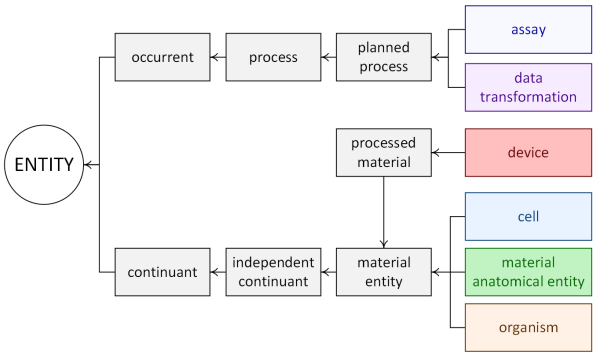

Fig. 1 Classes that describe the data within OBI. Six main classes that serve as descriptors are denoted with different colors. Arrow indicates *is a* property.

This work was funded by NIH NLM and NIDCD R01DC000354-S1.

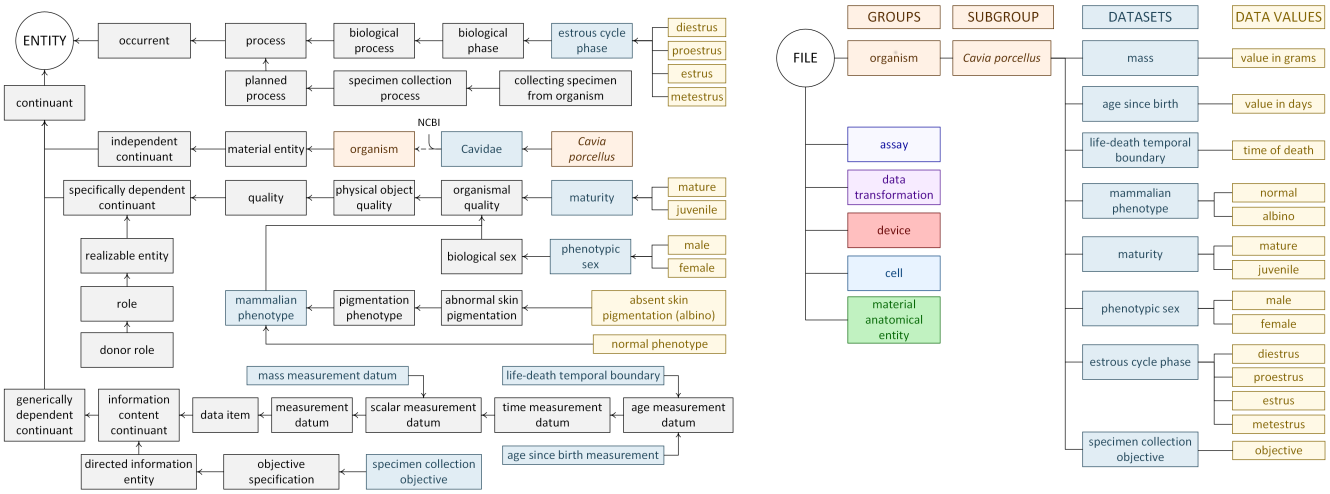

Fig. 2. Left panel. Classes that describe the organism arm of the data. They were imported from phenotype and trait ontology, PATO, mammalian phenotype ontology, MP, gene ontology, GO, ontology of biological attributes, OBA and NCBI. Dashed line indicates not all NCBI classes are shown. Arrow indicates: *is a* property. Right Panel. Data architecture of organism arm with other main Groups also shown.

analysis of data. Having examined other ontologies, we decided that Ontology for Biomedical Investigations, OBI [5] was the most robust and logical match for the data. However, it does not address all the concepts required for this data set. Approximately 100 are not found within OBI. We approached such cases by importing the terms from other ontologies, and by creating new classes (~ 20). These imported and new terms are grafted at suitable junctions onto a variant of OBI we have produced. Although these terms are being considered for acceptance into the OBI ontology, we stress that our goal is not to expand the ontology for its own sake, but to expand and use this application ontology to describe the data.

The six main classes used to define the data in OBI are shown in Fig. 1. The data are described with one electrophysiology assay: whole cell patch-clamp voltage clamp assay, with the devices (e.g. patch-clamp device) needed to perform the assay in a separate class. The assay was performed with outer hair cells isolated from cochlea of guinea pigs; hence cell, organism, and anatomical entity were the three other material entities. To extract key parameters from the electrical measurements we performed analysis and hence data transformation is the 6th class. These six main classes are the major arms of the data. In Fig. 2 we show the class structure for the organism arm with the associated data values that are normally recorded during an experiment. For example: the maturity and sex of the animal; whether the guinea pig exhibited a normal or albino phenotype; the mass of the animal; the age since birth; and, for adult females, the estrous cycle phase. Similar class relationships are formulated for the other five arms (data not shown).

IV. Design Data Architecture Based Upon Ontology

The original data were stored in MATLAB (Mathworks, MA), as a *struct*. MATLAB is proprietary software and not suitable for expansive data sharing, as this software must be licensed at a significant cost, and is not accessible to everyone. We transformed this data to Hierarchical Data Format version 5 (HDF5) [6] which was developed for storage of large and/or complex sets of data. We chose it because it is an open source format, with available open source viewers, it has conversion and editing application programming interfaces (APIs) for a variety of languages (including MATLAB), it supports complex and large data structures, and it already enjoys significant scientific usage.

The six main classes become the main branches of the data design (Fig. 2, right panel). All data associated with an individual outer hair cell were arranged in a tree configuration and saved to a file. The tree associated with the organism Group (terminology used with HDF5) is shown where the classes of the ontology now readily become the name for subgroups, datasets (terminology also used by HDF5), and data values. Similar mapping is performed for the other arms (Fig. 1) that describe the data.

Acknowledgment

We thank Anita Bandrowski, Owen Ellard, Ronna Hertzano, Barbara Jones, Elena Pormal, Joseph Santos-Sacchi, Jeffrey Teeters and Randy Vita for their contributions to this work.

References

1. Freedman, L.P., I.M. Cockburn, and T.S. Simcoe, *The Economics of Reproducibility in Preclinical Research.* PLoS Biol, 2015. **13**(6): p. e1002165.
2. Corbitt, C., et al., *Tonotopic relationships reveal the charge density varies along the lateral wall of outer hair cells.* Biophys J, 2012. **102**(12): p. 2715-24.
3. Hinard, V., et al., *ICEPO: the ion channel electrophysiology ontology.* Database (Oxford), 2016. **2016**.
4. Cook, D.L., et al., *Ontology of physics for biology: representing physical dependencies as a basis for biological processes.* J Biomed Semantics, 2013. **4**(1): p. 41.
5. Bandrowski, A., et al., *The Ontology for Biomedical Investigations.* PLoS One, 2016. **11**(4): p. e0154556.
6. *The HDF Group. Hierarchical Data Format, version 5 1997-2018* https://www.hdfgroup.org/HDF5/