



1

NeuroElectro: A Window to the World's Neuron Electrophysiology Data

Shreejoy J. Tripathy^{1,2,3,4,*}, Judith Savitskaya^{1,6}, Shawn D. Burton^{1,2}, Nathaniel N. Urban^{1,2}, and Richard C. Gerkin^{1,2,5,*}

¹Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA

²Center for the Neural Basis of Cognition, Pittsburgh, PA, USA

³Present Address: Centre for High-Throughput Biology, University of British Columbia, BC, Canada

⁴Present Address: Department of Psychiatry, University of British Columbia, BC, Canada

⁵Present Address: School of Life Sciences, Arizona State University, Tempe, AZ, USA

⁶Present Address: Graduate Program in Bioengineering, University of California, Berkeley and University of California, San Francisco, CA, USA

Correspondence*:

Shreejoy J Tripathy

177 Michael Smith Laboratories, 2185 East Mall, University of British Columbia, BC, Canada, V6T 1Z4, stripat3@gmail.com

Richard C Gerkin

School of Life Sciences PO Box 874501 Arizona State University, Tempe, AZ, 85287-4501, rgerkin@asu.edu

Neuroinformatics Infrastructure

2 ABSTRACT

The behavior of neural circuits is determined largely by the electrophysiological properties of the neurons they contain. Understanding the relationships of these properties requires the ability to first identify and catalog each property. However, information about such properties is largely locked away in decades of closed-access journal articles with heterogeneous conventions for reporting results, making it difficult to utilize the underlying data. We solve this problem through the NeuroElectro project: a Python library, RESTful API, and web application (at <http://neuroelectro.org>) for the extraction, visualization, and summarization of published data on neurons' electrophysiological properties. Information is organized both by neuron type (using neuron definitions provided by NeuroLex) and by electrophysiological property (using a newly developed ontology). We describe the techniques and challenges associated with the automated extraction of tabular electrophysiological data and methodological metadata from journal articles. We further discuss strategies for how to best combine, normalize and organize data across these heterogeneous sources. NeuroElectro is a valuable resource for experimental physiologists looking to supplement their own data, for computational modelers looking to constrain their model parameters, and for theoreticians searching for undiscovered relationships among neurons and their properties.

Keywords: neuroinformatics, electrophysiology, database, text-mining, metadata, API, machine learning, natural language processing

1 INTRODUCTION

21 Brains achieve efficient function through implementing a division of labor, in which different types of
22 neurons serve distinct functional and computational roles. One striking way in which neuron types differ
23 is in their electrophysiology properties. Though the electrophysiology of many neuron types has been
24 previously characterized and documented across decades of research, these data exist across thousands of
25 journal articles, making cross-study neuron-to-neuron comparisons difficult.

26 Neurophysiology lacks a centralized resource where consensus data on basic physiological
27 measurements from many neuron types and studies are accessible for reference and subsequent meta-
28 analyses. For example, though it is common for neurophysiologists to measure and report neuronal
29 measurements such as resting membrane potential and input resistance, there is not a public database
30 which compiles this information. In other domains of neuroscience such efforts have made more progress.
31 In the domain of neuroanatomical connectivity, information on connectivity between different brain
32 regions is being compiled by experts at the Brain Architecture Management System project (BAMS)
33 across hundreds of publications (?). Parallel to this effort is the WhiteText Project, which addresses a
34 complementary goal by algorithmically mining brain region connectivity statements from journal abstracts
35 using biomedical natural language processing (bioNLP) methods (??). Similarly, in the domain of
36 neuroimaging, the NeuroSynth Project has mined fMRI-based brain activation maps from published x,y,z
37 coordinate data tables from thousands of neuroimaging publications (?). These literature-based methods
38 can be contrasted with projects such as NeuroMorpho.org (?) and ModelDB (??), which index neuron
39 morphological reconstructions and computational models for simulating neuron activity by obtaining this
40 information directly from investigators.

41 Success among these projects can be defined according to different criteria. Such criteria include
42 completeness and comprehensiveness; for example, what percentage of relevant connectivity studies
43 are indexed within BAMS? How many different neuron types are contained within the NeuroMorpho
44 database? Alternatively, success can be defined in terms of the utility of these databases in driving
45 subsequent research, like the use of BAMS as a resource for discovering relationships between brain
46 region connectivity and gene expression (?) or the use of NeuroMorpho to discover general scaling
47 relationships among the morphology of neuron types (?). Similarly, NeuroSynth is widely used by
48 cognitive scientists as a starting point for designing functional imaging studies. Thus while these projects
49 are not yet comprehensive and likely contain data records of varying quality, these resources may
50 nevertheless be employed to draw novel inferences.

51 These projects are logically divided according to their methods for obtaining the source data: through
52 the use of manual methods like expert curation or user contributions versus automated methods such
53 as text-mining. Notably, these approaches differ in their scale and accuracy; while algorithmic methods
54 can “scale-up” and be applied to arbitrary numbers of publications, they typically have a lower accuracy
55 relative to human-curated content (?). This lower accuracy is often attributed to the rich lexical complexity
56 of biomedical texts which often require considerable context and background knowledge to understand
57 and parse (??). The competing constraints of scale versus accuracy pose a challenge for large-scale
58 compilation of neuroscientific data.

59 Here, we built a custom infrastructure framework for extracting electrophysiological measurements for
60 specific neuron types from published neurophysiology articles. These measurements included properties
61 such as input resistance and resting membrane potential, as well as associated metadata (i.e., article-
62 specific methodological details). Our methods combine algorithmic literature text-mining, drawing from
63 the approach used by NeuroSynth (?) where neurophysiological measurements are primarily extracted
64 from data tables, as well as manual curation, leveraging the background knowledge of domain experts.
65 The resulting neurophysiology database, named NeuroElectro, can be interactively viewed and explored
66 through a public web interface at <http://neuroelectro.org>.

2 MATERIALS, METHODS, & RESULTS

2.1 OVERVIEW

67 We describe and validate our semi-automated methodology for obtaining neuronal biophysical
68 measurements directly from published reports in the literature (summarized in Fig. 1). After obtaining
69 full article texts from publishers, we then used text-mining algorithms to identify concepts specific to
70 electrophysiology and neuron types, which we then validated manually.

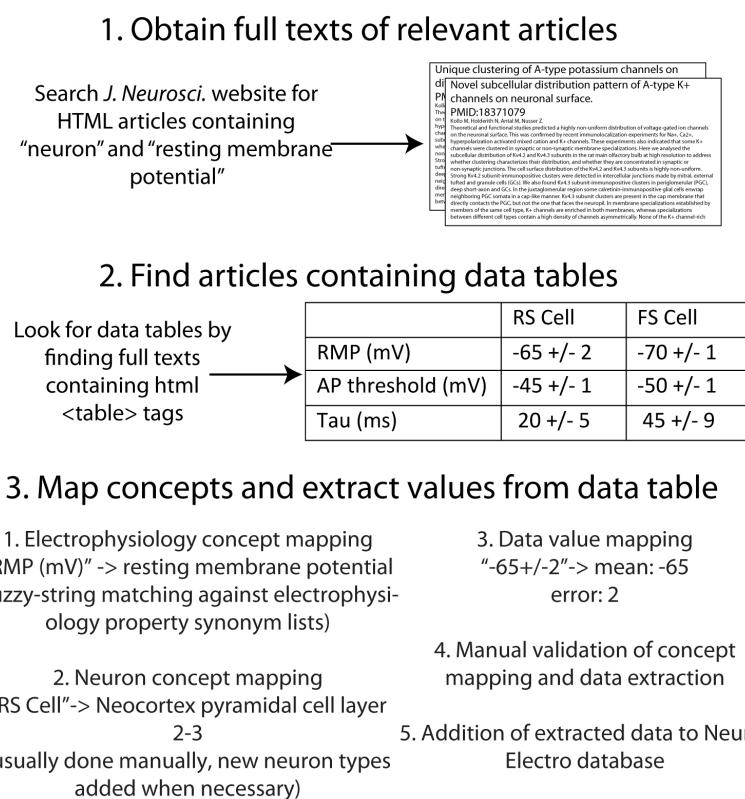


Figure 1. Illustration of workflow for obtaining electrophysiological information from the research literature.

2.2 ARTICLE IDENTIFICATION

71 We obtained electrophysiological data from 10 neuroscience specific journals (Table 1), which include:
72 *Journal of Neuroscience*, *Journal of Neurophysiology*, and *Journal of Physiology* (among others). We
73 selected these journals because they often devote a significant fraction of an article's main text, tables,
74 and figures to detailed characterizations and summaries of intrinsic neuronal biophysical properties.

We obtained tens of thousands of potentially relevant full article texts directly from publisher websites. We first identified potential articles that were likely to contain information relevant to neuron biophysics using the native search functions provided within the journal websites and only downloaded articles containing a specific list of terms including “input resistance” and “resting membrane potential” (Fig. 1). This pre-selection step allowed us to identify and download only articles that contained data relevant to our project. Upon identifying candidate articles, we then downloaded the full text of each potentially-relevant article as HTML; articles downloaded from the publisher Elsevier (e.g., *Neuron* and *Brain Research*) were downloaded as XML using the provided text-mining API and subsequently converted

Table 1. Statistics of journals represented in the NeuroElectro database. Listing of journals and counts of articles downloaded (articles obtained), articles with published data tables containing neurophysiological information which has been manually validated by an expert curator (validated), and articles which likely contain information in a data table which has not yet been manually curated (not validated). Not validated articles are those which have at least 4 algorithmically assigned electrophysiological concepts within a data table.

Journal	Articles obtained	Validated	Not validated
J. Neurosci.	19002	104	560
J. Neurophysiol.	12078	94	555
J. Physiol. (Lond.)	10543	44	235
Neuroscience	3035	14	205
Eur. J. Neurosci.	2495	7	117
Brain Res.	3017	7	146
Neuron	1657	4	43
Epilepsia	463	2	23
Neurosci. Lett.	1468	2	34
Hippocampus	208	2	10

[

83 to HTML. We chose to work with HTML (as opposed to PDF or XML) because HTML provides
 84 a machine-readable markup of the article's content, allowing us to easily identify relevant elements
 85 within the article – such as data tables and the methods section – using publicly available HTML-
 86 parsing tools (here we used the Beautiful Soup HTML-processing library implemented in Python:
 87 <http://www.crummy.com/software/BeautifulSoup/bs4/doc/>). Furthermore, because
 88 HTML is a single semi-structured standard used across publishers, we could write relatively generic
 89 HTML-processing algorithms applicable to content published across journals. However, our focus on
 90 using HTML limits us to relatively newer articles – typically those published after 1997 – because most
 91 publications before this time are only available as scanned PDF files.

92 We stored the HTML-enhanced full text of each article in our database and associated each article
 93 with its corresponding PubMed ID (<http://www.ncbi.nlm.nih.gov>). These 8-digit IDs serve as
 94 publisher-independent unique identifiers for each article, and allow us to use PubMed-specific tools, such
 95 as a powerful API (i.e., PubMed eutils, <http://www.ncbi.nlm.nih.gov/books/NBK25500/>).
 96 For example, this API provides the ability to query each article's MeSH terms (MEdical Subject Headings)
 97 and returns basic methodological information such as animal species and strain.

2.3 ELECTROPHYSIOLOGICAL PROPERTY IDENTIFICATION

98 2.3.1 *Rationale for focusing on electrophysiological property extraction from data tables.* In order to
 99 algorithmically extract information on neuron electrophysiology from these articles, we needed to first
 100 specify the data types of interest. Our preference was to obtain as much detailed information about neuron
 101 electrophysiological properties as possible: ideally, this would include raw data corresponding to recorded
 102 electrophysiological traces. In mining information from articles, we were presented with multiple options
 103 (illustrated in Fig. 2), including extraction from: 1) the text of the article including figure captions, 2)
 104 the figures of the article, or 3) data tables presented within the article. In addition to these, authors often
 105 submit supplemental materials and figures which also contain neurophysiological data.

106 Given the challenges in mining raw electrophysiological traces from figure images, we instead focused
 107 on obtaining information about basic neuronal electrophysiological properties, such as input resistances
 108 and resting membrane potentials. Though this information is often presented within the text of the article,
 109 it is usually presented in complex sentence structures that are difficult to accurately parse algorithmically.

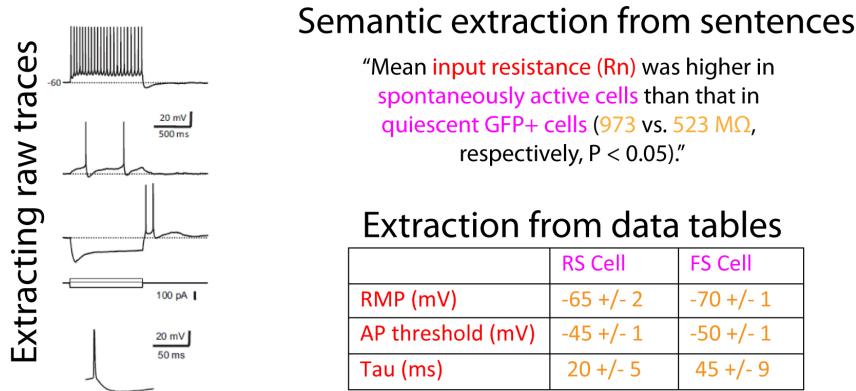


Figure 2. Illustration of the sources within an article containing information relevant to neuron electrophysiological properties. Data on neuronal electrophysiological properties are presented within article figures and raw traces, sentences within the article text, and formatted data tables. The raw traces and references are from ? and the data table is a constructed example. Colored text indicates electrophysiological concepts (red), neuron concepts (pink), or neurophysiological data (yellow).

110 Published data tables, on the other hand, present a unique opportunity for electrophysiological data
 111 extraction, since common techniques exist for extracting information from structured tables (?). Moreover,
 112 because tables succinctly summarize multiple attributes of a collected dataset, the effort of an expert
 113 curator can be put to best use when validating tables relative to validating content mined from article
 114 sentences or figure panels. While we estimate that only 5-10% of electrophysiology articles contain
 115 data tables, there is sufficient redundancy within the field (i.e., multiple investigators often publish
 116 articles on the same neuron type) that focusing on data tables nevertheless yields substantial coverage
 117 of electrophysiological properties across nearly all major neuron types.

118 **2.3.2 Extracting information on electrophysiological properties.** In extracting electrophysiological data
 119 we took advantage of the fact that certain measurements are commonly made during intracellular
 120 recordings. For example, such recordings are commonly used to: 1) measure a neuron's resting
 121 membrane potential, 2) to apply hyperpolarizing current injections for measurement of input resistance
 122 and membrane time constant, and 3) to apply depolarizing current steps to evoke action potentials (spikes)
 123 and enable measurement of characteristics such as spike threshold, width, and amplitude.

124 We developed an electrophysiological lexicon comprising 28 measurements that we found to be
 125 commonly reported in the literature, largely based on previously published definitions (??). To account
 126 for subtle differences in terminology that authors use to refer to the same electrophysiological concept
 127 (e.g., resting membrane potential is often referred to as "rmp" and " V_{rest} "), we also identified a
 128 common list of synonyms to map to each concept. Together, these electrophysiological concepts and
 129 their synonyms define a preliminary ontology for electrophysiological concepts (included in Supplemental
 130 Materials). Moreover, this physiological measurement ontology can serve as a scaffolding for a more in-
 131 depth ontology of electrophysiological investigations (e.g., Ontology for Experimental Neurophysiology,
 132 ?). The terms in our preliminary ontology are also indexed and defined within NeuroLex (<http://neurolex.org>, (?)).

A

Table 1.
Comparison of electrophysiological properties in adult +/+ and stg/stg in deep layer cortical neurons

	+/+	stg/stg	P
V_r , mV	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)	ns
R_{in} , MΩ	170 ± 25 (20)	170 ± 13 (22)	ns
Time constant, ms	26.9 ± 2.6 (14)	32.1 ± 3.1 (22)	ns
AP overshoot, mV	37.0 ± 3.7	34.1 ± 3.14 [11–58] (23)	ns

B

Table 1.
Comparison of electrophysiological properties in adult +/+ and stg/stg in deep layer cortical neurons

	+/ <i>Concept: Neocortex pyramidal cell layer 5-6</i>	stg/stg	P
V_r , mV <i>Concept: resting membrane potential</i>	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)	ns
R_{in} , MΩ <i>Concept: input resistance</i>	170 ± 25 (20)	170 ± 13 (22)	ns
Time constant, ms <i>Concept: membrane time constant</i>	26.9 ± 2.6 (14)	32.1 ± 3.1 (22)	ns
AP overshoot, mV <i>Concept: spike overshoot</i>	37.0 ± 3.7	34.1 ± 3.14 [11–58] (23)	ns

Figure 3. Example data table illustrating mark-up and annotation of entities. A. Example published data table containing neurophysiological information. Data table from ?. B. Same as A, but semantically marked up with algorithmic and manually curated annotations. Markups in red and pink indicate electrophysiological and neuron type concepts and yellow indicates extracted data measurements. Note that here the textual string “+/+” and “stg/stg” refers to the normotypic and manipulated condition, respectively.

134 To identify data corresponding to electrophysiological properties reported within a data table,
 135 we developed algorithms to search data table header elements and assess whether these elements
 136 corresponded to any of the electrophysiological concept synonyms in our ontology. We first identified
 137 table header elements by searching for table elements composed primarily of non-numeric characters.
 138 For each putative header element, we then used fuzzy string matching algorithms (implemented using
 139 the fuzzywuzzy library in Python: <https://github.com/seatgeek/fuzzywuzzy>), to assess
 140 the textual match between the header element and each of the electrophysiological synonyms. These
 141 fuzzy matching algorithms combine a number of string match metrics into a single “match value”,
 142 including whether a pair of strings completely match, contain matching substrings, or contain matching
 143 but misordered substrings. If the table header and electrophysiological synonym match value exceeded
 144 a specified threshold, the table header and corresponding row or column of numeric values were
 145 automatically mapped to the electrophysiological concept. Similarly, we mapped whole rows or columns
 146 to specific neuron types recorded during normotypic or “wild-type” conditions.

147 We then manually corrected cases where these algorithms misassigned an electrophysiological concept.
 148 For example, a common algorithmic mis-assignment was the case when an author used the string “EPSP
 149 amplitude” to refer to the electrophysiological concept excitatory post-synaptic potential amplitude. In
 150 these cases, our algorithms incorrectly mapped this string to “spike amplitude” because the former concept
 151 is not in our current ontology. In a test sample of 279 articles that were manually curated, we found that
 152 78% of concept-matchings (901/1152) were identified correctly with no supervision, with the remainder
 153 manually corrected.

154 2.3.3 Accounting for differences in electrophysiological definitions across investigators. By focusing
 155 on textually matching the electrophysiological terms in each table to a list of electrophysiological
 156 concepts, we are implicitly assuming that electrophysiological properties are measured in the same way
 157 by investigators across different articles. For example, the most common method that electrophysiologists

use to measure a neuron's spike properties is to record from the neuron in current-clamp mode and apply peri-threshold depolarizing currents to evoke 1-2 spikes over several hundred milliseconds or more. The neuron's spike amplitude is then commonly measured by calculating the difference between the neuron's voltage at spike threshold and spike peak for the first evoked spike (e.g. (??)). However, experimental differences exist between how investigators measure and compute these properties; we divide these differences into roughly 3 categories: *protocol*, *calculation*, and *condition* differences. For example, investigators can use different experimental protocols to measure the spike amplitude, like evoking spikes using current steps much greater than rheobase current required to elicit a single spike (*protocol differences*). Additionally, the spike amplitude itself can be calculated in different ways, such as using the neuron's resting membrane potential as the baseline instead of the spike threshold (*calculation differences*). Furthermore, the value of spike amplitude that an investigator reports will also be affected by specific experimental conditions such as the animal species or age and recording solution temperature or contents (*condition differences*).

When manually curating the text-mined content for the most commonly reported electrophysiological properties (resting membrane potential, input resistance, membrane time constant, spike half-width, spike amplitude, and spike threshold), we took care to account for and remove cases where the investigator had calculated an electrophysiological measurement using an inconsistent methodology (e.g., protocol or calculation differences). However we note that we could not identify all of these cases (in particular: spike amplitude, input resistance, and membrane time constant), in part because investigators did not always explicitly define how these measurements were calculated within their article. We note that in cases where we pool measurements which are measured using inconsistent protocols or calculations, this will tend to add unexplained variance to our data set. Given these measurement inconsistencies, we provide our recommendations for how these electrophysiological properties should be reported in future investigations via our electrophysiology ontology (see Supplemental Materials).

2.4 NEURON TYPE IDENTIFICATION

2.4.1 *Using neuron types defined by NeuroLex.* To extract physiological information specific to individual neuron types, we had to identify which neuron types were reported in each article. However, in many cases uniquely identifying the neuron type(s) reported in any given study and mapping these to a canonical “neuron type” is difficult. This difficulty arises in part because investigators use different criteria for classifying neurons, including electrophysiological, morphological, or molecular characteristics (???).

To define canonical neuron types, we chose to use an existing list of approximately 250 neuron types and definitions provided by NeuroLex, a community-sourced, expert-defined collection of neuron types (<http://neurolex.org>; ???). Moreover, we chose to use NeuroLex to keep our database consistent with existing resources and to enable future researchers to combine these resources seamlessly. NeuroLex also provides synonyms for each neuron type, which we utilized to identify the neuron type(s) in each article. In cases where a neuron type was investigated in the literature across multiple articles but not indexed within NeuroLex (e.g. cerebellar nucleus neurons), we manually added this neuron type to our database’s listing and provided this neuron type to the NeuroLex neuron curators for incorporation (Gordon Shepherd, personal communication). Our specific criteria for identifying each of the neuron types reflected in the database are given in the Supplemental Materials.

2.4.2 *Identifying specific neuron types within an article.* Because of the complexity in unambiguously identifying neuron types, we used a mixed text-mining and manual approach to map the neuron types studied in each article to canonical NeuroLex neuron types. First, we used text-mining algorithms to provide an initial “best guess” of the most likely neuron type. Specifically, we used a bag-of-words approach (?) on the full article text. This approach ignores the serial structure of the words in the document and utilizes only the frequency of occurrence of each word within the document. We next compared the article’s word-frequency histogram to the listing of neuron synonyms provided by NeuroLex, ranking all neuron types by their likelihood of being actually studied within that article. In comparison to articles that

we manually curated, we found that this automated approach accurately identified the neurons studied in each article with an accuracy of 30% (120 of 399 total) and up to 55% when defining success as the studied neuron appearing as one of the top three neuron types suggested by the bag-of-words method. Because of the relatively low accuracy of an automated-only approach, we added a manual curation step where a curator identified the recorded neuron type using HTML drop down menus enriched by the bag-of-words search (e.g., Fig. 4). As previously described, we mapped individual data table elements and corresponding rows or columns to specific neuron types recorded under normotypic conditions. We note that currently we only identify data from normotypic or “control” neurons represented in tables, but plan to identify data from additional conditions in future work (e.g. from pharmacologically manipulated or genetically modified animals).

2.5 EXTRACTION OF ELECTROPHYSIOLOGICAL DATA VALUES

After identifying specific electrophysiological properties and neuron types reported in a data table (corresponding to row or column table headers), we then algorithmically extracted the data corresponding to the table intersection of these (Fig. 3). We developed custom string regular expressions (?) to parse the string corresponding to the numeric data. Specifically, we found that data strings were often of the form: “XX ± YY (ZZ)”, where XX, YY, and ZZ refer to the mean, error term, and sample size (i.e. the “n”), respectively. Often the number of replicates or error measurement were not reported or were reported in alternative ways within the table. Presently, the error term is not resolved as either a standard deviation or standard error measurement in the current version of NeuroElectro, but could easily be resolved in future iterations.

When designing our processing algorithms, we parsed data strings from right to left: first searching for data entities contained within parentheses, then for entities contained to the right of the ± term, and finally the remaining term which we assumed to refer to the mean term. We found that occasionally data were reported as “XX (LL - HH)” – where LL and HH indicate the lower and upper limits of a data range – and accounted for these cases similarly. We used regular expressions to identify entities such as digits, decimal signs, parentheses, and ± signs. We then converted the individual data elements which were encoded as textual strings of digits to double precision decimal entities before storing these into our database. Our focus here was primarily on parsing the data record mean value (i.e., summarizing the properties of a number of recorded neurons), but also extracted and stored the error term and sample size where possible. Using these methods, we were able to extract 2176 electrophysiological values for 93 distinct neuron types within 279 articles.

2.6 MANUAL VALIDATION OF AUTOMATED DATA EXTRACTION

Following these automated concept identification and data extraction steps, we manually validated associated concepts and fixed incorrect concept mappings as necessary. We developed custom-HTML and javascript code to allow human curators to graphically interact with downloaded HTML data tables and “mark-up” entities within the table (Fig. 4). This code allows for textual based elements of the HTML table to be semantically annotated using drop down menus and text fields. Moreover, because annotation is implemented via user interfaces composed of interactive web pages and drop down menus, these user interfaces are simple enough to be utilized by other expert curators with little formal instruction.

2.7 METADATA IDENTIFICATION

Given the strong relationships between experimental conditions, such as animal species or recording temperature, and electrophysiological measurements (e.g., input resistances are known to decrease when measured in neurons from older animals (??)), we also identified information on article-specific experimental conditions by extracting this information primarily from each article’s methods section. For each article, we found the methods section by developing custom HTML tag filters for each journal (e.g., common publisher-defined HTML tags for methods sections are “Methods” or “Experimental

Table 2. A partial listing of metadata attributes and extraction methodology. Metadata attributes are extracted through combining PubMed Medical Subject Heading terms (MeSH Terms) and custom regular expressions (Regex). Regular expression column (or MeSH Term column) indicates specific regular expressions (or MeSH terms) used for identifying metadata concept entities.

Metadata Concept	Values	Extraction method	Regular Expression	MeSH Term
Species	Rats Mice	MeSH Term only		Rats Mice
Electrode Type	Guinea Pigs Patch-clamp sharp	MeSH Term + Regex	"whole cell" or "patch clamp" "sharp electrode"	Guinea Pigs Patch-Clamp Techniques
Animal Strain	Fischer 344 Long-Evans Sprague-Dawley Wistar C57BL BALB C	MeSH Term only		Rats, Inbred F344 Rats, Long-Evans Rats, Sprague-Dawley Rats, Wistar Mice, Inbred C57BL Mice, Inbred BALB C
Prep Type	in vitro in vivo cell culture model	MeSH Term+Regex	"slice" or "in vitro" "in vivo" "culture" "model"	Cell Culture Techniques Computer Simulation
Jxn Potential	Not Corrected Corrected	Regex	"not junction potential" "junction potential"	
Rec Temperature	Continuous value Room temperature	Regex	"record ... C" or "experiment C" "record room temperature"	
Animal Age	Continuous value	Regex	Find digits near: "P#-#" or "P#-P#"	

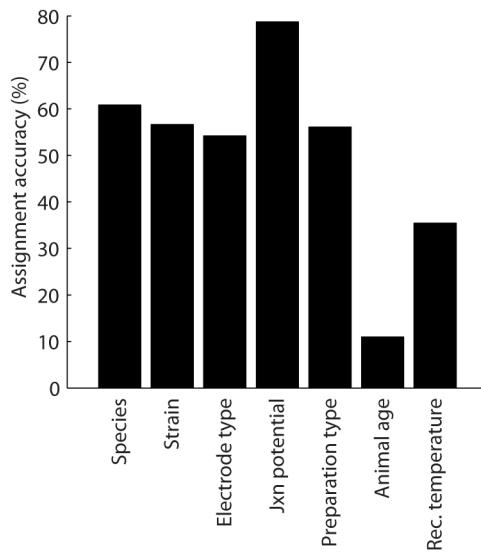
248 procedures"). For each metadata entity that we focused on (species, animal strain, electrode type, preparation type, liquid junction potential correction, animal age, recording temperature), we devised 249 custom automated text searching methods to identify these based on combining regular expressions 250 (?) with PubMed MeSH terms (Table 2). In other words, rather than taking a machine-learning based 251 approach and training classifiers (?), we took a rule-based approach and developed custom rules for 252 identifying metadata entities. For example, to identify whether the recording electrode's liquid junction 253 potential was corrected for in the study (?), we searched for whether the character string "junction 254 potential" was mentioned within the methods section and, if so, whether the sentence or phrase containing 255 the term was explicitly negated (indicating that the junction potential was not corrected for). Here, we 256 identified and parsed distinct sentences within the methods section using tools provided within the Natural 257 Language Tool Kit in Python (?).

259 Following automated identification of article metadata, we then manually checked each article to 260 ascertain that algorithmically-tagged metadata was identified correctly and, as before, we corrected 261 misidentified content as necessary through the use of custom HTML forms. We found that the mean 262 accuracy of algorithmic metadata assignment was approximately 50% (Fig. 5) and was typically lower 263 for identifying continuous-valued metadata (e.g. animal age or recording temperature) relative to nominal 264 metadata such as species and electrode type.

Table 1.

Comparison of electrophysiological properties in adult +/+ and stg/stg in deep layer cortical neurons

	+/ Concept: Neocortex pyramidal cell layer 5-6	stg/stg
V _r , mV Concept: resting membrane potential	-74.4 ± 1.5 (25)	-73.7 ± 1.7 (27)
R _{in} , MΩ Concept: input resistance		
•Ephys Prop Neuron		
'input resistance' <input type="button" value="Submit"/>	170 ± 25 (20)	170 ± 13 (22)
'None selected' 'cell capacitance' 'input resistance'		
'resting membrane potential' 'membrane time constant' 'spike amplitude' 'spike half-width' 'spike threshold' 'rheobase' 'firing frequency' 'AHP duration' 'cell diameter'	26.9 ± 2.6 (14) 37.0 ± 3.7 [12–67] (14)	32.1 ± 3.1 (22) 34.1 ± 3.14 [11–58] (23) 6.9 ± 1.1 (14) 1.5 ± 0.7 (12)

Figure 4. Example of human validation of algorithmically assigned content. All textual elements of a table are enhanced using HTML and javascript to allow for assignment of neuron or electrophysiological concepts using drop down menus. Example data table from ?.**Figure 5.** Accuracy of metadata assignment using automated methods alone.

2.8 OBJECT MODELS AND RELATIONAL DATABASE

265 We stored extracted data and metadata using a relational database implemented in MySQL (<http://dev.mysql.com/doc/refman/5.6/en/>) built from a Python Django object model (<https://www.djangoproject.com/>). The object model contains classes for a number of fields, such as full
 266 article texts, electrophysiological properties, neuron types, synonyms, electrophysiological data values,
 267 and experimental metadata (Fig. 6). A useful feature of the relational nature of the database is that it
 268 enables linking between classes (e.g., linking between neuron types and electrophysiological properties
 269 reported by a single investigator across multiple articles). This linking feature facilitates efficient and
 270 arbitrary querying of data; for example, querying for known electrophysiological data on olfactory bulb
 271 mitral cells recorded *in vitro* and published between the dates 2000 and 2004. For example, such a feature
 272 could be used to assess whether the properties of olfactory bulb mitral cells have changed as a function of
 273 time or are dependent upon whether the data are collected *in vitro* or *in vivo*.
 274

2.9 WEB APPLICATION

276 The primary results of NeuroElectro are viewable at <http://www.neuroelectro.org> where the
277 data can be interactively explored.

278 2.9.1 *Human interface* The web interface is organized around neuron types and electrophysiological
279 properties. For example, each neuron type has its own webpage where extracted data corresponding
280 to specific electrophysiological properties is graphically and interactively displayed (graphical plot
281 interactivity implemented using the jqPlot javascript toolbox, <http://www.jqplot.com/>). Users
282 can thus visualize the mean and variability of electrophysiological values across papers, view references
283 plus experimental metadata, and easily navigate to primary data from specific papers. Furthermore, users
284 can view electrophysiological data across all of the neuron types in the database - putting phenotypic
285 properties of a given neuron type into the larger context of other neuron types located throughout the
286 nervous system.

287 The web application also contains preliminary features to allow website visitors to contribute to the
288 NeuroElectro resource. For example, users can suggest articles which contain electrophysiological data
289 which are not already in the database. We also invite visitors to become “expert curators” for neurons
290 of interest. In the future, we plan to build functionality that will allow investigators to upload raw and
291 summary data, such as recorded voltage and current traces. In addition, we plan to continue mining the
292 literature and adding neurophysiological measurements as they are published.

293 2.9.2 *API* An initial API (application programmer interface) providing public access to the
294 electrophysiological data is described at <http://neuroelectro.org/api/docs/>. This RESTful
295 API allows contents of the NeuroElectro database to be dynamically retrieved in JSON or XML format to
296 utilize this data within external applications. For example, using the current API, a developer could build
297 an application which dynamically queries NeuroElectro for all data corresponding to layer 2/3 neocortical
298 pyramidal cells and then uses this data to constrain parameters for a Hodgkin-Huxley type neuron model
299 (?). Example use cases of the current API (version 1) include:

- 300 • <http://neuroelectro.org/api/1/n/> : Returns a list of all neurons with electrophysiological
301 data indexed in NeuroElectro.
- 302 • <http://neuroelectro.org/api/1/nedm/?nlex=sao830368389> : Returns a list of all
303 indexed data on CA1 pyramidal cells (queried using the NeuroLex identifier for CA1 pyramidal cells,
304 sao830368389).
- 305 • http://neuroelectro.org/api/1/nes/?e_name=Input\%20Resistance\&n__name=Cerebellum\%20Purkinje\%20Cell : Returns a data record composed of the
306 mean, standard deviation, and n, summarizing input resistance measurements from cerebellar
307 Purkinje cells based on all indexed articles in NeuroElectro database. Here the database query is
308 performed using the textual strings for the electrophysiological and neuron type concepts.

310 Our future plans are to work with domain ontologists to further develop the existing API into a formal
311 relational data format (RDF) specification, allowing further querying and extending of NeuroElectro into
312 additional resources.

3 DISCUSSION

313 We have developed, applied, and validated a methodology and pipeline for extracting – from existing
314 literature on cellular neurophysiology – measurements of basic biophysical properties from diverse neuron
315 types throughout the nervous system. Currently, the NeuroElectro database contains 2344 manually

316 curated electrophysiological measurements from 98 neuron types from 335 publications. Of these
317 electrophysiological measurements, 2176 (93%) were obtained from 279 (83%) publications using the
318 semi-automated approach described here. In addition, we machine-extracted and manually validated 1667
319 methodological conditions (metadata) from these publications. This represents the single largest collection
320 of neurophysiological data ever compiled and represents a potentially valuable tool for scientific discovery.

3.1 SPECIFIC BENEFITS PROVIDED BY THE SEMI-AUTOMATED APPROACH

321 One of the key advantages of the approach described here is that the automated pipeline identifies
322 publications which are likely to contain content relevant to our domain area (i.e., measurements of
323 neuronal biophysics). Thus a human needs only to manually curate the content first identified by the
324 algorithms as being likely relevant, instead of having to identify the relevant content *de novo*. Moreover,
325 the automated identification of neuron types in articles allows us to target manual curation efforts
326 to publications likely to contain data from specific neuron types, such as neurons that are currently
327 underrepresented in the database.

328 Given our laboratory's focus on olfactory circuits, we conducted a natural experiment to compare
329 the efficacy of biophysical property extraction using these semi-automated methods versus traditional
330 methods which do not make use of algorithmic text-mining as a pre-processing step. In a seven-
331 hour curation session, a senior graduate student in our laboratory identified 91 electrophysiological
332 measurements (focusing on resting membrane potential, input resistance, membrane time constant, spike
333 amplitude, spike width, and spike threshold) from 35 articles for 7 olfactory bulb neuron types using only
334 prior knowledge of which articles and investigators were likely to have reported such electrophysiological
335 data. In a comparable seven-hour curation session using our semi-automated methods, a single curator
336 (with similar expertise to the first curator) identified 551 electrophysiological measurements from 70
337 articles across 40 neuron types throughout the nervous system. Moreover, this comparison would likely
338 tilt even more in favor of the semi-automated methods had the curators been less familiar with the primary
339 literature.

3.2 SCALABILITY OF CURRENT APPROACH

340 We note that multiple steps in our approach require manual intervention by an expert curator in order for
341 electrophysiological measurements to be extracted with an acceptably low error rate. Namely, an expert
342 curator needs to specify which neuron types are recorded from in each article and where data from the
343 normotypic or "control" states of these neurons are textually referenced within a data table. Moreover,
344 given the current accuracy of the unsupervised algorithmic assignment of electrophysiological concepts
345 and experimental metadata (78% and 50%, respectively), these also need to be manually validated and
346 corrected as required by an expert. Given the necessity of these manual steps, the scalability of our current
347 approach is limited by our ability to manually curate this information or by our ability to improve the
348 error rate of the automated methods. Despite this limitation, our current pipeline is still much faster
349 than a purely manual one. The methodology could be further improved by addressing falsely matching
350 entities (such as EPSP amplitude in section 2.3.2) by either auto-suggesting alternative matches returned
351 by string-similarity algorithms or by simply adding these concepts to the electrophysiolgical ontology.
352 Moreover, these improvements would facilitate formally computing the sensitivity and specificity of these
353 entity recognition methods.

3.3 EXTENSIONS AND IMPROVEMENTS TO THE CURRENT SEMI-AUTOMATED ALGORITHMS

354 Currently, neuron type identification is a critical bottleneck in our approach. One potential improvement
355 would be to replace the nonspecific bag-of-words approach we are currently using in favor of a
356 bioNLP classifier-based approach (?). Specifically, we propose adapting the named entity recognition
357 methodology used by the WhiteText project for tagging brain regions mentioned in literature (??) and

358 first identifying spans of text likely to pertain to a neuron type before mapping these textual spans to a
359 individual neuron type within the neuron ontology. Moreover, such an approach could easily incorporate
360 common neuron type acronyms and abbreviations (?).

361 The approach described here is highly effective for extracting biophysical measurements presented
362 within machine-readable data tables published within journal articles. However, the current requirement
363 that these data tables exist in a machine parseable format, such as HTML or XML, limits this approach
364 from being directly applied to older manuscripts, which are only available as scanned images. Existing
365 approaches, such as optical character recognition technology (OCR; e.g. (?)) may be applied toward this
366 problem in the future.

367 Given the relatively low accuracy of the automated approach to identifying neuron types, there may
368 be several avenues through which this process can be improved. For example, we note the automated
369 approach was particularly ineffective when the neuron type investigated within an article was not already
370 described in NeuroLex or when the neuron had an insufficient list of synonyms associated with it. The
371 current implementation of NeuroElectro also does not consider common neuron type acronyms (e.g., that
372 olfactory bulb mitral cells are commonly referred to as “MCs”). Adding acronym identification to future
373 iterations will thus likely improve the automated approach (??). Moreover, our current implementation of
374 the bag-of-words algorithm would likely be improved via minor improvements, such as only identifying
375 neurons using the text of the abstract or results and discarding text from the introduction or discussion. As
376 neuron identification forms the major bottleneck in the scalability of NeuroElectro due to the requirement
377 for manual curation, we plan to address this bottleneck in future revisions.

3.4 FUTURE METHODS FOR DATA EXTRACTION

378 A more pressing issue with the current approach is its focus on extraction from data tables. We estimate
379 that only 5-10% of published electrophysiological data is contained within tables, while the remaining
380 90-95% is presented within article text or figure images. Given our preference to obtain data in their
381 most raw form, we initially considered extraction of data from figures, e.g. voltage traces of neuronal
382 activity. However, digitizing article figures (presented by publishers as images) into a form that can
383 be further analyzed presents multiple challenges. Though techniques and tools exist to digitize figures,
384 substantial amounts of manual effort are required to employ them correctly, making this figure-based
385 approach difficult to scale to increasing numbers of articles without also employing a large team of
386 human curators. While automatically extracting measurements from figure images will likely prove
387 challenging, our methods can likely be adapted to operate on article text, perhaps by making use of
388 bioNLP methodologies currently used for relationship extraction in the identification of connected brain
389 regions (?) or interacting pairs of proteins (?).

390 However, more widespread inclusion of fundamental electrophysiological measurements in data
391 tables in future publications will strongly facilitate data-sharing and meta-analysis initiatives such as
392 NeuroElectro. We believe that, if successful, the use of NeuroElectro will influence the practices of
393 scientists writing papers and reporting results. Specifically we believe (and hope) that it will result in
394 scientists reporting more basic physiological data overall as well as reporting more data in machine-
395 parsable tables. This change would make it easier for scientists to find and make use of data collected
396 by others. Moreover, such a culture shift has the potential to make science function more effectively and
397 efficiently to facilitate discovery.

DISCLOSURE/CONFLICT-OF-INTEREST STATEMENT

398 The authors declare that the research was conducted in the absence of any commercial or financial
399 relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGEMENT

400 We thank William Cohen, Gordon Shepherd, and Renaud Richardet for discussions and comments on
401 the manuscript. We are especially grateful to all of the investigators whose collected data are represented
402 within the NeuroElectro database. We thank the academic journal publishers (in particular, Elsevier and
403 Wiley and Highwire) for allowing us access to their full-texts for text-mining. We are especially grateful
404 to all of the investigators whose collected data are represented within the NeuroElectro database.

405 *Funding:* This work was supported by a National Science Foundation Graduate Research Fellowship
406 and a R. K. Mellon Foundation Fellowship (to S.J.T.), an Achievement Rewards for College Scientists
407 Foundation Fellowship and NIDCD NRSA F31DC013490 (to S.D.B.), NIDCD award F32DC010535 and
408 NIMH award R01MH081905 (in support of R.C.G.), and NIDCD award R01DC005798 (to N.N.U.).

SUPPLEMENTAL DATA

