Implementing Effective Biocuration Process, Training, and Quality Management Protocols on Undergraduate Biocuration of Amyotrophic Lateral Sclerosis

An Undergraduate Thesis
Presented to
The Academic Faculty

by

Rachel Wilcox True

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Science in Biomedical Engineering

Georgia Institute of Technology
May 2015

Implementing Effective Biocuration Process, Training, and Quality Management Protocols on Undergraduate Biocuration of Amyotrophic Lateral Sclerosis

Approved by:

Dr. Cassie S. Mitchell, Advisor
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Robert Lee
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology

Dr. Joseph Le Doux
Wallace H. Coulter Department of Biomedical Engineering
Georgia Institute of Technology

TABLE OF CONTENTS

ABSTRACT

Biocuration is manual scientific collection, annotation and validation of literary information of biological and model organisms into a single database. Successful biocuration processes involve those with an extensive collection of literature, a user-friendly database interfaces for entering and analyzing data from published papers, and highly regulated training and quality assurance protocols. Due to the rapid expansion of biomedical literature, an efficient and accurate biocuration process has become more valuable due to the magnitude of data available in published literature. As the biocuration process incorporates undergraduates, it is critical that the medium for data collection is simple, ergonomic, and infallible. A reconstructed FileMaker Pro database was introduced to previously trained undergraduate students for process evaluation. Streamlining the biocuration process and grouping data structure to be more intuitive were two goals the new database interface hoped to achieve. The creation of a rigorous training program and strict quality management protocol is needed to prepare the lab for the introduction of efficient biocuration processes. Through the database designing process, training protocols were drafted to effectively call the biocurator's attention to important changes in the interface design. Upon prototyping the database, entry errors were reviewed, training protocols were adjusted, and the quality protocols were drafted. When the combination of undergraduate biocurators and the reconstructed database under these new protocols was compared to statistics in the biocuration field, results proved to show increase in both productivity rates as well as accuracy rates. By having such efficiency at the undergraduate level, subject matter experts will no longer be required to perform this type of research and can focus on analysis. This will increase research productivity and reduce costs in the overall biocuration process. With over 12,000 published papers regarding Amyotrophic Lateral Sclerosis on Pubmed in 2014 alone, this revolutionary combination could lead to quickly finding a suitable cure for these patients.

LITERATURE REVIEW

Biomedical researchers are making greater use of biocuration, the manual extraction of biological data within published literature, to more rapidly draw conclusions on the effects of different drugs and treatments on complex diseases such as Amyotrophic Lateral Sclerosis (ALS), a fatal disease characterized by the loss of function of motor neurons (Silani et al., 2010). This cell death occurs in the spinal cord and leads to muscle deterioration, eventual paralysis, and respiratory failure (Zhang et al., 2014). Animal model databases that focus on the underlying pathophysiology of diseases such as ALS, including cellular through systemic level mechanisms, provide an untapped resource. For example, a single PubMed search on ALS transgenic mice returns over 2,600 articles. In 2014 alone, nearly 1500 articles were published specifically on ALS, and the cumulative total literature base for this pathology exceeds 12,000 articles, on PubMed alone. One might think that with this wealth of information we would have ALS well in hand. Yet a life-extending treatment still is not available. ALS remains one of the most obstinate neurological diseases with no answers, only intertwined experimental observations recorded through a vast amount of individual publications. Furthermore, at this point, there is a lack of ability to pinpoint the underlying cause of ALS. The ability to biocurate and perform bioinformatics analysis on this data set could provide critical understanding of the etiology of ALS. In this study, a database of the SOD1 G93A Mouse Model of ALS is used, the most published ALS transgenic mouse model.

Many current databases provide researchers with detailed information in two categories; animal genetics or clinical epidemiology (Drabkin et al., 2012; Mandler et al., 2001). The primary focus of animal model databases is genomics, genetics, and biological data of various animals of concern. For example, Mouse Genome Informatics (MGI) integrates information about sequencing data, phenotypes, embryonic expression, and more, all pulled from peer-reviewed literature (Drabkin et al., 2012). Another, ParkDB, is a relational database for observing Parkinson's Disease in mice, zebrafish, and humans (Taccioli et al., 2011). Finally, clinical databases, such as the ALS Patient Care Database, focus on documenting patient medical history, patient background, and disease statistics (Mandler et al., 2001).

INTRODUCTION

Biocuration is the manual scientific collection, annotation and validation of literary information of biological and model organisms into a single database. Using previously published literature, biocurators aim to collect, organize and share valuable information to promote accessibility. Biocuration therefore lends itself to indirect collaboration, enhancing research processes and leading to scientific findings that might otherwise not be readily apparent. As the amount of published literature continues to grow, effective biocuration processes become invaluable (Lu et al., 2012). However, this is an immensely time-consuming process that requires painstaking attention to detail (Hirschman et al., 2012; Bandrowski et al., 2012). The ability to pull and compile data, to pick and choose what is pertinent to the investigator, creates opportunities for rapid advancements in all biological fields. Because biocuration makes such a large impact, biocurators need an efficient method to capture data from publications (Howe et al., 2008). Plenty of researchers have created their own databases in hopes of improving the difficulties associated with biocuration.

Current databases are designed from the standpoint of the biocurator or the bioinformaticist. That is, they are developed in order to record pre-declared relevant information, such as gene mutation or cell lines. What current databases fail to do, causing the biocuration process to be more difficult, is focus on the intuitive breakdown of information provided in each piece of literature. The database used in this study successfully merges the two, creating an ergonomic user interface fit for undergraduate level biocurators. By directing the focus of the database on article structural components (e.g. abstract, figure caption, etc.), undergraduates are able to cohesively recapture both qualitative and quantitative data without sacrificing data collection rates of skilled professionals. Biocuration processes are typically tedious, time consuming, and performed by skilled professionals in the field. Database user interfaces are typically complex, counterintuitive and often filled with unnecessary systemic loops causing the biocuration process to be painstakingly challenging. A combination of trained undergraduate biocurators and a user-friendly database interface not only optimizes the data extraction process, it allows more rapid strides in the discovery of a potential treatment or cure for ALS.

The goal of this project is to develop database platform training, biocuration, and quality control protocols that will successfully enable scientific article biocuration by undergraduate research associates. We hypothesize that under these protocols, scientific journal article data can be rapidly curated by undergraduates while maintaining higher than 97% accuracy, a rate that is comparable to professional PhD biocurators. Within the scope of this study, this lab hopes to refine the database platform, refine the biocuration protocol, develop a biocurator training protocol, develop a quality control protocol, and assess biocuration accuracy and productivity using the developed platforms and protocols. By optimizing this entire process, the bioinformatics field will have great potential in curing ALS.

METHODS

In order to optimize the undergraduate biocuration rates in this lab, there needed to be a combination of a user-friendly database platform, proper biocuration, training, and quality management protocols, strict assessments of associate performance, and proper filtering of literature. With over 12,000 published papers on ALS, the first step was to narrow the search terms.

Inclusion Criteria:

In this study, a database of the SOD1 G93A Mouse Model of ALS is used, the most published ALS transgenic mouse model. This experimental model inspects the pathophysiology of ALS, including *in vitro*, *in vivo*, and mixed type experiments of transgenic mice. Literature search criteria "G93A" OR "Mouse" AND "ALS" OR "Amyotrophic Lateral Sclerosis" was searched in all fields on PubMed to collect the papers' publishing information. As of October 2014, the publishing information of 3412 papers has been saved to the database. Further exclusion of papers is conducted during the biocuration process, if the paper is deemed out of the inclusion criteria.

Database Platform:

To allow for more efficient analysis, an intuitive database was created to reflect the data structure established by the common structure in G93A literature as opposed to the pre-imposed ontology with a specific type of research in mind. Non-quantifiable information was used by bioinformaticists and biocurators for the data-driven flexible ontology, labeling for comparative analysis, and to scope out holes in the field of research. The majority of the quantifiable data is contained in the figures or tables (Burge et al., 2012). After assessing the information viable for bioinformatics analysis, the FileMaker Pro 13 database interface prototype was created. (Roussey et al., 2011). Filemaker Pro Server Console is employed to maintain the database on the internal server, enabling data entry and data access from multiple different computers simultaneously. Data structure was broken down and categorized into four entry layouts; Paper, Figures, Data Series, and Response Values, to streamline the biocuration process. All papers from the search criteria above were imported from EndNote, automatically filling the Paper layout. The remaining layouts separate data to be extrapolated from the paper into intuitive groups, eliminating wasted time in traversing the piece of literature and decrease data-entry error. Although the interfaces call specific information that may be difficult to extract, scroll cues and hint boxes are available for every field to assist the biocurators.

Biocuration Process:

The biocuration process is broken into predetermined levels that vary in the depth of captured information. Depending on associates' familiarity with the database, responsibility,

previous biocuration performance, and desired data capture style, each level of biocuration differs in procedure. Regardless of what level the associate is working at, they are expected to produce work at a productivity rate of 17 entries per hour with an accuracy rate greater than 97%. As associates excel in their work, they are introduced to the next level. As outlined in Table 1, data capture techniques coincide with the breakdown of data structure outlined in the database. The field definitions were divided up to match the respective biocuration level and given to the associates prior to their training date.

Level 1:

To begin, associates are given primary resources into understanding the SOD1 G93A Mouse Model as well as FileMaker Pro to read up on. The main responsibility of Level 1 associates is document capturing. Associates must retrieve and save the pdf of the desired article using Pubmed searches through online resources at Georgia Institute of Technology as well as Emory University.

Level 2:

Once associates have adequate understanding of the background of this lab and have shown proficiency in Level 1, they advance to Level 2. The main responsibility of Level 2 associates is figure caption capturing. Associates are responsible for extracting information from the figure and table caption text from documents captured by Level 1 associates and entering it into the database into the respective categories. Level 2 associates primarily work in the Figure layout of the database.

Level 3:

Once associates have adequate understanding of how to correctly extract data and have shown proficiency in Level 2, they advance to Level 3. The main responsibility of Level 3 associates is data recapturing. Associates are responsible for extracting quantifiable primary experimental data from figures captured by Level 2 associates and entering it into the database into the respective categories. Level 3 associates primarily work in the Data Series and Response Value layout of the database. When a paper has successfully transcribed by all levels of associates, it is considered a curated article, in which case the Level 3 associate deems it fully curated and it is removed from the search terms of associates.

Training Protocol:

Each level of the biocuration is challenging not only in extracting the correct information but inputting the information into the database as well. Furthermore, with multiple biocurators retrieving comparable information, challenges may arise in sorting and organizing information in the database. Therefore, a descriptive and structured protocol must be implemented to ensure a regulated procedure that produces comparable results among all associates. Due to the update in data structure breakdown, training manuals and procedures were needed to reflect and coincide with the levels of the biocuration process. To begin, the lab strived to find associates from a variety of undergraduate specializations including, but not limited to, biomedical

engineering, computer science, and industrial engineering. Approximately 22 associates are chosen as biocurators each semester for an appropriate balance of workers and computer resources available. The training protocol will ensure a baseline standard data collection method amongst the associates. Training sessions are conducted as interactive workshops with approximately a 5:1 associate to manager ratio that last approximately two hours. This interactive environment allows for hands on experience with the database and an open opportunity for the associates to ask questions. Coinciding with the levels of biocuration, training is broken into stages encompassing a variety of learning styles, allowing for maximum procedure retention and assurance of proper data collection. The stages or steps, as outlined in Figure 1, include interactive workshops, individual reading and comprehension, and assessments

Level 1:

The intention of training a Level 1 associate is to ensure they are familiar with the process of biocuration and the fundamentals of the disease pathology. Prior to the training session, associates are given a primary list of definitions and background information they are expected to review. During the training session, associates are taken through a step-by-step procedure of finding and saving a sample paper, intended to ensure they have access to the Georgia Institute of Technology and Emory University systems and provide them with hands on practice.

Level 2:

The intention of training a Level 2 associate is to ensure the correct methods of text mining are used at a broad Figure level. Beginning with Level 2 training, each associate is given their own clone of the database to act as a drawing board for their learning while maintaining the integrity of the full database's information. Prior to the training session, associates are expected to review over field definitions pertaining to Level 2 work. During the training session, the quality management team walks through a sample paper for demonstration of how to properly perform figure capture, providing difficult examples and clear explanations of what is expected for each field entry. At the end of the training session, associates are subject to the Figure Capture Competency Assessment.

Level 3:

The intention of training a Level 3 associate is to ensure the correct methods of text mining are used at a specific data recapture level. For Level 3 training, associates are to work in their clone databases given to them during Level 2 training. Prior to the training session, associates are expected to review over field definitions pertaining to Level 3 work. During the training session, the quality management team walks through a sample paper for demonstration of how to properly perform data recapture, providing difficult examples and clear explanations of what is expected for each field entry. At the end of the training session, associates are subject to the Data Recapture Competency Assessment.

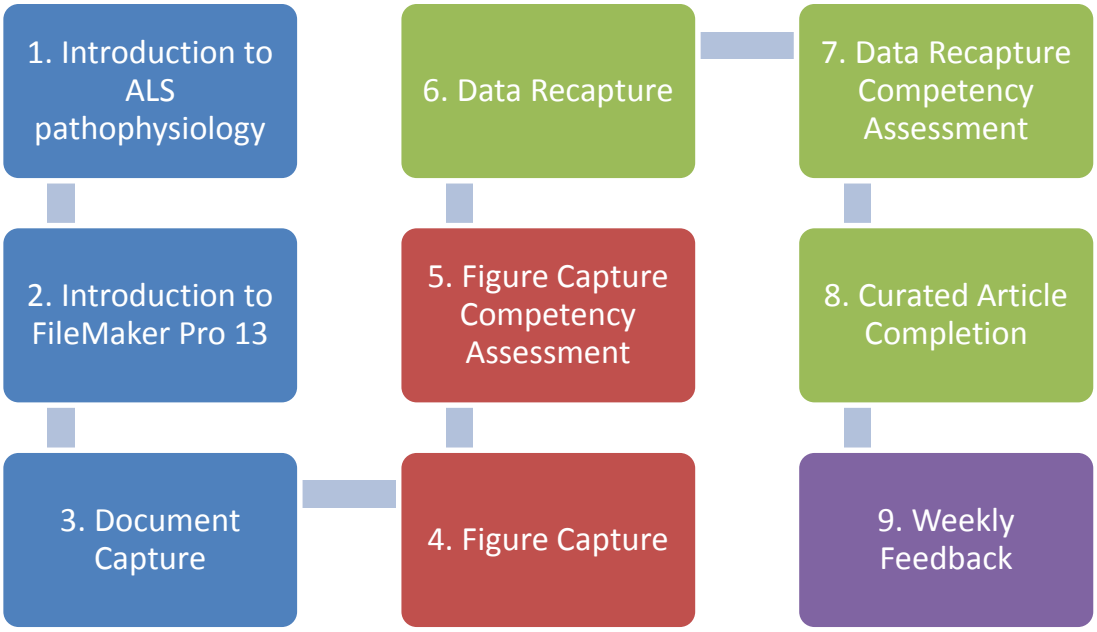


Figure 1. Training Outline. Steps taken by quality management team to ensure standard teachings across all associates. Blue represents Level 1. Red represents Level 2. Green represents Level 3. Purple represents other responsibilities of the quality management team.

Competency Assessments:

Upon completion of training at each level, an exam is given where associates are to biocurate assigned papers into their respective clone databases for assessment of their understanding during training. Exam is meant assess each associates competency in the given level of biocuration. Through the assessments, the lab managers are also able to infer if alterations to protocols must be made, common errors are present, database updates need to be made, and more, allowing for an ease of accessing and using the database. Furthermore, assessment can be used to do the following; compare accuracy to alternative methods of biocuration, compare productivity to alternative methods of biocuration, provide the biocurator with feedback for ease of usage. Because 100% of entries are reviewed for quality management during training, the assessment is rigorous and very strict in the passing criteria. Associates are given feedback on their performance and expected to make the given changes to their database entries for assurance of complete understanding of the process. Pending passing remarks on the training exam, the associates are either granted access to the prototype database or asked to perform a retraining.

Quality Control Protocol

In order to maintain and ensure correct biocuration processes occur, the development of a strict quality management system was created. The biocuration process is time consuming and requires highly accurate data collection to ensure effective collaboration of information. The quality management team, comprised of experienced students with at least two semesters of biocuration experience, is responsible for ensuring highly productive and accurate data collection, benchmarked by lab specific minimum acceptance values of productivity and accuracy, and weekly feedback to the associates. The quality management team independently assesses each biocurator at a 7-10:1 associate to manager ratio, beginning with the training process. Standards created during prototype testing were expected to be withheld in all areas. Failure to achieve either standard led to infraction points, the chosen method of grading for the lab. Accumulation of infraction points meant a drop in letter grade for the associate. All calculations and evaluations were recording in a password protected excel file for easy management, allowing the quality management to see trends in associates work as well as quickly sum and analyze numbers.

Productivity:

Productivity rates were monitored by the management team via automated timestamps of entries and scribe IDs, unique associate identifiers used for quality management. With the help of FileMaker scripts, the management team summed the hours of recorded work by the associate and the number of entries created within that time. Associates were subjected to withhold a minimum of 17 entries per working hour in the prototype database interface, a standard created during prototype testing.

Accuracy:

Accuracy rates were determined by the ratio of number of errors made to total entries made per associate and weighted by severity and types. During training and the first few weeks after training, 100% of associates' entries are checked to identify any errors. After that, entries are randomly selected and at least 50% of the entries are checked on a weekly basis. The maximum number of entries checked is a function of the associates' present and historical performance. Associates were subjected to withhold a minimum of 97% accuracy in the prototype database interface, another standard created during prototype testing.

Feedback:

When associates finish their work for the week and after their productivity and accuracy rates have been determined, the quality management team is responsible for providing each associate feedback on their progress and weekly work. Any errors found in the associates work is to be shared both for grading enlightenment as well as error corrections. Associates are expected to correct their mistakes as they are informed by the quality management team. Therefore, weekly feedback is intended to keep the database accurate and allow for each associate to grow in their biocuration abilities.

Protocol Assessment

At the end of a16-week semester determined by the Georgia Institute of Technology registrar, a compilation of lab averages of both productivity and accuracy rates were created by the quality management team. The same compilation was created from results of a previous semester, using old training protocols and the original database interface. The two compilations are compared with the intention of assessing the effectiveness of the new protocols on the two measurements taken by the quality management team.

RESULTS

Database Platform

The data in the G93A mouse model field can be best categorized into four distinct mouse type-treatment combination labels: “Wild Type Control”, “Wild Type + Treatment”, “G93A Control” and “G93A + Treatment.” These labels were used to categorize the mice into genetic background, the initial breakdown of information collected. Table 1 represents the resulting predetermined field definitions used to create the database interface to streamline the biocuration process. Figure 2 visually represents where a biocurator can find the various field definitions within a paper.

Table 1: Field Definitions

Figures	**Definition**
Accession ID	The foreign key for the Papers table.
Type	Type of the figure - Figure or Table.
Number	Number of the figure or table.
Panel	Panel letter of the table, if applicable.
Part	Separate parts of a figure panel, if applicable. Different plots are considered separate parts.
Title	Title of the figure or table. Usually the first sentence of the figure caption
Description	Description of the figure or table. Usually the figure caption except the first sentence.
Quantifiable	Whether the figure or table is quantifiable (convertible to numeric). A quantifiable figure typically includes a bar graph, line graph and scatter plot with a trendline.
Experiment Type	One of the following 3: **In vivo:** An experiment conducted on living mice. A common measure includes rotarod performance, body weight, cumulative survival, etc. **In vitro:** An experiment conducted outside a living mouse, such as a cell line, tissue section, etc. A common measure includes intracellular ion concentration, cell viability, etc.Mixed: An experiment that started with a living mouse but where in vitro assessments were made. Examples include experiments on blood cells extracted from living mice, evaluation of motor neuron count in a tissue extracted from a sacrificed mouse.
FigTab ID	The primary key for each figure entry
Data Series	**Definition**
FigTab ID	The foreign key for the Figures table
Mouse Type	Type of the mouse breed. G93A, Wild Type, G85R, etc.
Genetic Background	The genetic background of the mouse. Mostly C57BL/6 or B6SJL.
Genotype	Zygosity of the mouse line. Heterozygous, Homizygous or Hemizygous.
Sample Size	Number of samples or trials used in the experiment.
Treatment	Any external treatment or reagents applied on mice or samples, where applicable.
Note	Any extra information about the treatment, mouse type, etc.
Copy Number	Amount of copies of transgenic genes in transgenic mice. Either High or Low
Time of Onset	Time of onset observed in an in vivo experiment.
Time of Death	Time of death observed in an in vivo experiment.
Data Series ID	The primary key for each entry.
Response Values	**Definition**
Data Series ID	The foreign key for the Data Series table.
Time point	The time at which the measurement was taken.
Response Description	The thing that is measured. Usually the Y-axis on a graph.
Value	Quantified data point from the figure or table.
From Ontology	Ontology of what is tested or is supposed to cause changes in mice. Usually treatment.
To Ontology	Ontology of what changes in mice. Usually outcome.
Response Value ID	The primary key for each entry.

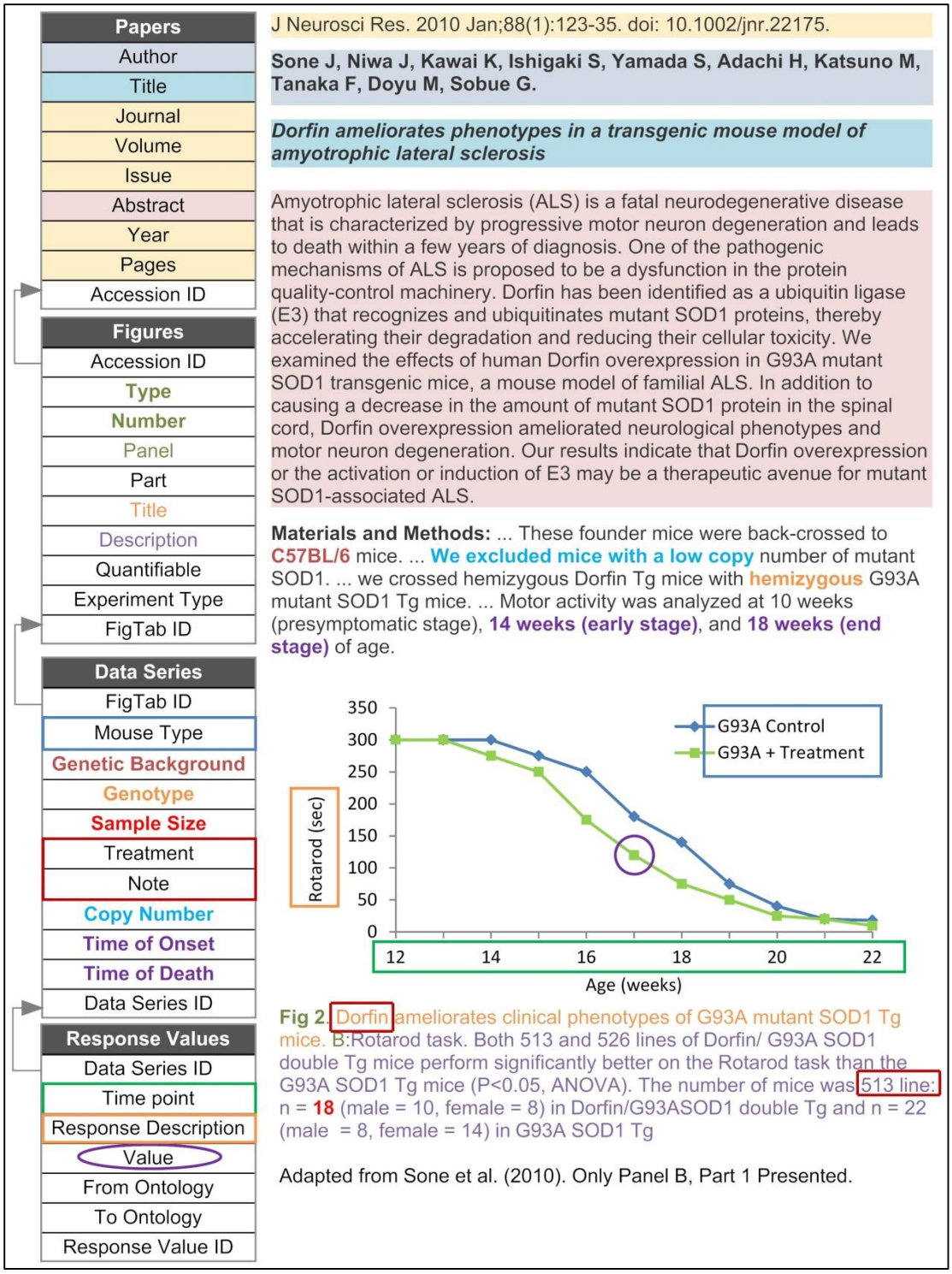


Figure 2. Biocuration Paper Overview. Visual representation of where each field definition can be extracted from a paper. Format of the field definitions on the left match the respective data from the sample paper on the right.

Biocuration Process

After assessing and organizing the field definitions, the biocuration process was divided into the levels, as described in Table 2, for ease of training, understanding, and intuitive paper breakdown. These productivity techniques aided in the creation of the training and quality management protocols. This breakdown reduces the number of scrolls through the paper one associate must perform.

Table 2: Capture Type Definitions

Level	Productivity Technique	Definition
Level 1	Document capture	Obtaining a pdf download of the full manuscript text.
Level 2	Figure caption capture	The scraping of figure and table caption text from the captured document and entering it into the database.
Level 3	Data recapture	The process of extracting quantifiable primary experimental data from the figures and tables of the captured document.
Total	Curated article	An article that has been fully transcribed in the database, including document capture, figure capture and data recapture

Training Protocol

Training protocol, as seen in Appendix 1, is descriptive and broken into stages coinciding with the levels of biocuration. Broken into nine easy steps, as detailed in Table 3, the associate and quality manager have clear processes to follow.

Table 3. Training steps and brief descriptions.

Step	Detail
Introduction to pathophysiology	Associates are given brief primer and experimental overview of the ALS G93A transgenic mouse model, including some review articles to read
Introduction to FileMaker Pro 13	Associates are shown how to access the database and navigate between different entry tables in the database
Document Capture	Associates are shown how to access, download, and appropriately archive full-text pdf articles
Figure Capture	Associates are shown how to distinguish between quantifiable and non-quantifiable figures (i.e. immunostaining pictures) and how to appropriately divide the figures and tables into panels and parts.
Figure Capture Competency Assessment	Associates are given 5 articles to successfully perform document and figure capture into clone database. Must have >95% accuracy to move to the next step.
Data Recapture	Associates are shown how to A.) determine if the data to be recaptured meets the curation inclusion criteria; B.) capture necessary experimental methods information that is not in the figures (i.e. transgenic mouse strain, definition and time of ALS onset, definition and average time of experimental endpoint); C.) identify a data series (follow a trendline for wild type versus G93A, etc.); and D.) estimate quantitative values from a figure
Data Recapture Competency Assessment	Associates are given 5 articles to successfully perform document capture, figure capture, and data recapture into clone database. Must have >95% accuracy to move to the next step
Biocuration Release	Associates are granted access to the main database
Weekly Feedback	Quality management team assesses weekly work for productivity and accuracy rates

Quality Control Protocol

The quality management team aimed to calculate each associates' productivity and accuracy rate each week. The calculations were then inputted into a master excel document for record keeping. Embedded in this document are functions called to sum and/or average the intended output. The final two columns of the excel document, seen in Figure 3, assess if the associate has hit both the productivity and accuracy requirements or if they have earned an infraction point. This excel document is up kept weekly by the quality management team when they perform the weekly feedback.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Week 4 (9/6 - 9/12) (Sat-Fri)															
2	Scribe ID	Data Entry Hrs	Data Entry Type	Prod Req (entries/ hr)	Corr Req (%)	Papers	Figures/Tables	Data Series	Response Values	Total Entries	Total Entries/ Hrs	Errors	Entries Checked	% Correct	Productivity Infraction	Accuracy Infraction
3	AJ	2	standard	17	97.5	1	48	70	85	204	102	3	204	98.5	No	No
4	BIM	2	standard	17	97.5	0	21	72	69	162	81	2	162	98.8	No	No
5	CSH	4	standard	17	97.5	0	52	69	61	182	46	4	182	97.8	No	No
6	CWI	4	standard	17	97.5	0	53	71	69	193	48	4	193	97.9	No	No
7	EAM	2	standard	17	97.5	0	46	96	69	211	106	3	211	98.6	No	No
8	GMB	4	standard	17	97.5	0	46	80	90	216	54	2	216	99.1	No	No
9	IMV	4	standard	17	97.5	0	49	76	70	195	49	1	195	99.5	No	No
10	JPK	3	standard	17	97.5	0	49	74	72	195	65	3	195	98.5	No	No
11	KDB	4	standard	17	97.5	0	45	76	69	190	48	2	190	98.9	No	No
12	MLJ	4	standard	17	97.5	0	52	70	69	191	48	4	191	97.9	No	No
13	MTD	4	standard	17	97.5	0	37	51	62	150	38	4	150	97.3	No	Yes
14	RC	3	standard	17	97.5	0	38	41	70	149	50	3	149	98.0	No	No
15	SRP	4	standard	17	97.5	0	28	71	84	183	46	2	183	98.9	No	No
16	TEK	4	standard	17	97.5	0	41	56	70	167	42	1	167	99.4	No	No
17	TZ	4	standard	17	97.5	1	38	43	83	165	41	2	165	98.8	No	No
18	YS	4	standard	17	97.5	0	50	65	291	406	102	5	291	98.3	No	No
19	Standard Average			17.0	97.5	0.1	43.9	69.6	89.4	203.0	63.1	2.9	194.8	98.5		
20	Standard Standard Deviation			0.0	0.0	0.4	9.5	13.8	58.6	61.3	23.8	1.1	33.3	0.5		

Figure 3. Master Excel Document Sample. Master excel document used by the quality management team over the duration of the semester to assess associates performance long term.

Productivity:

FileMaker Pro 13's automated scripts were used to sum the productivity of each associate. Figure 4 represents the code used to produce the productivity script. This code calls the entire database in the respective fields searching for two main things, the scribe ID of associate under review and the time frame of the weekly feedback. By using automated timestamps when the associate creates and modifies an entry, the code sums the associates work. The quality management team must simply type in the associates scribe ID and the date range into the FileMaker fields, then the script does the rest of the work. The output of this script was then entered into the master excel document for record keeping. It was observed that over the span of the semester, productivity rates showed a general increasing trend through time, meaning that as the associates got more experience, they were able to work at faster paces, collecting more data in the same amount of time. This trend can be seen in Figure 5.

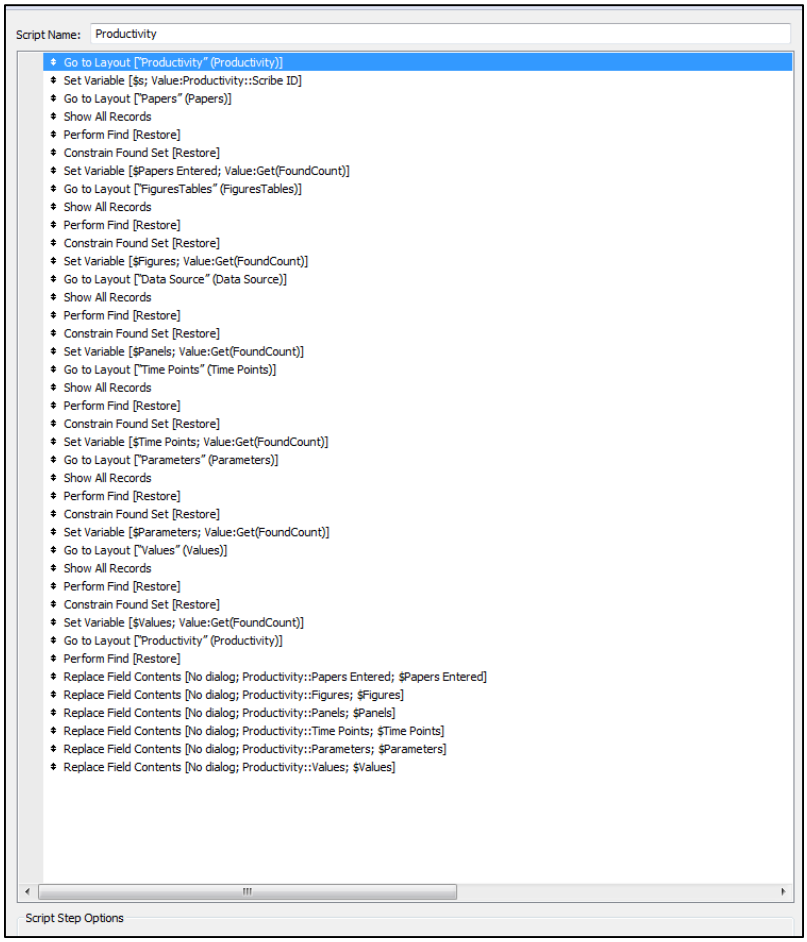


Figure 4. Productivity script code.

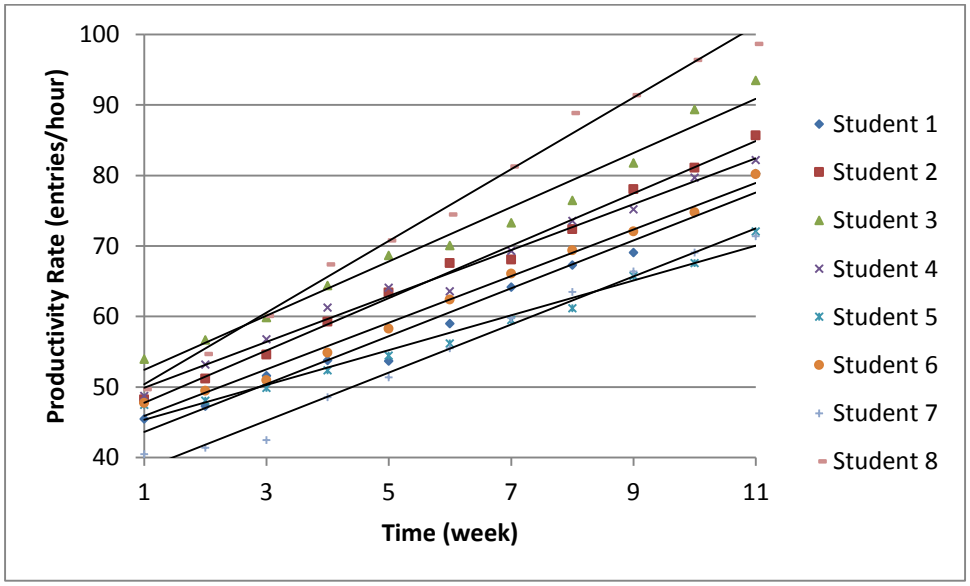


Figure 5. Productivity rates of 8 data recapture associates over the span of the semester. Extracted from the productivity script.

Accuracy:

After determining the productivity rate of each associate, the quality management team assessed the errors made each week by each associate. Using the entries called out in the productivity script, quality management team reviewed the associates work in search for errors and mistakes. As seen in Table 4, errors can be classified into various categories. Each mistake found is to be recorded and given to the associate for feedback. The sum of the errors made is then inputted into the master excel document for rate assessment. It was observed that over the span of the semester, the number of errors showed a general decreasing trend through time, meaning that as the associates got more experience, they were able to work more accurately by reducing the number of errors made. This trend can be seen in Figure 6. Furthermore, at the end of the semester, the total errors of each type were observed and summed, as seen in Figure 7, thus allowing the quality management team to note what the error trends were in order to adjust protocols accordingly for future semesters.

Table 4: Common Errors Definitions

Accuracy Type	Definition
Estimation error (EE)	Incorrectly estimated value from a figure or table incorrectly. Data recapture values entered in the database that deviated by more than 5% from the actual value in the figure were considered estimation errors
Partial data capture (PDC)	Failure to collect all data from a figure or table. For example, trendline data was recaptured only for the G93A mouse when the figure had trendline data for both the G93A mouse and the wild type mouse.
Mislabeling (ML)	Failure to enter correct values in the appropriate field. For example, time of onset is incorrectly entered in the time of death field.
Extraneous data (ED)	Data collected that does not meet curation inclusion criteria. For example, data for a G85R trendline is recaptured even though the curation project only includes G93A data recapture (see Inclusion Criteria).

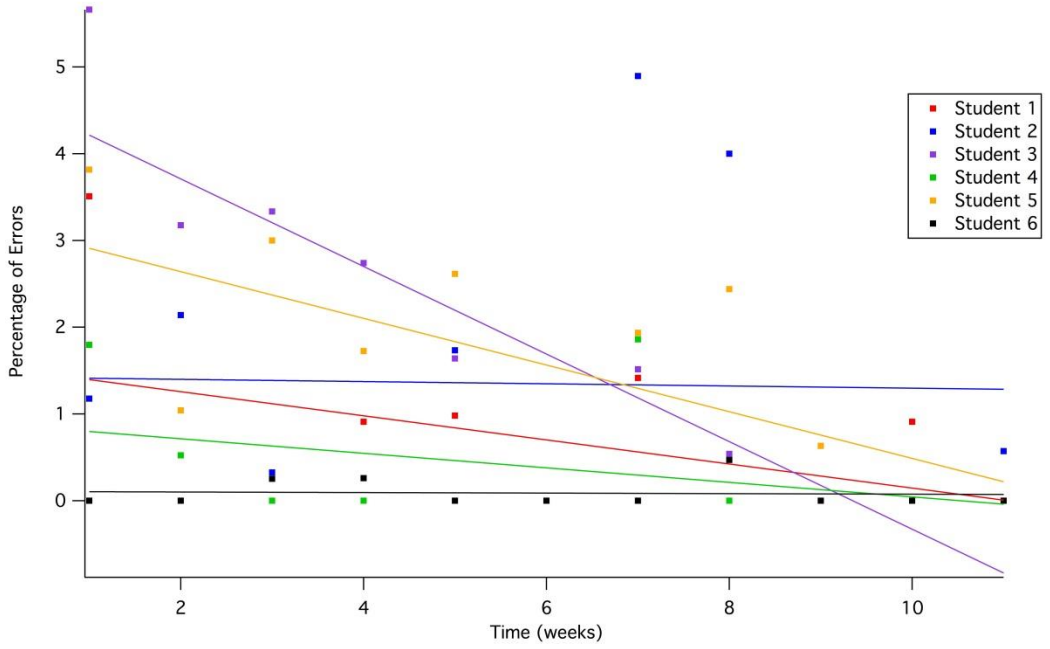


Figure 6. Errors of six data recapture associates over the span of a semester taken as a percentage of total entries per week.

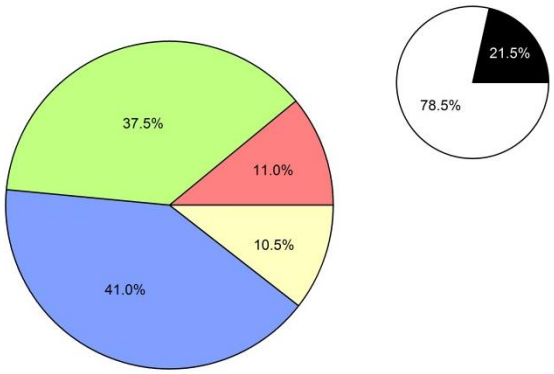


Figure 7. The percentage of each type of error of the total errors. Green represents partial data capture errors. Red represents estimation errors. Yellow represents errors of data. Blue represents mislabeling errors. The sub-pie graph shows the total critical (black) and non-critical (white) errors.

Protocol Assessment

When in comparison to database interfaces used in previous semesters, the reconstructed database interface prototype demonstrates and improvement in both of the tested criteria; productivity and accuracy rates. Twenty two undergraduate students at Georgia Institute of Technology tested the database interfaces over the course of a 16 week semester. The reconstructed prototype yielded greater productivity by approximately 5 entries per hour per associate while it yielded greater accuracy by approximately 3% per associate as seen in Figure 8. Both criteria were subjected to paired t-tests yielding significant differences for both, $p<0.05$.

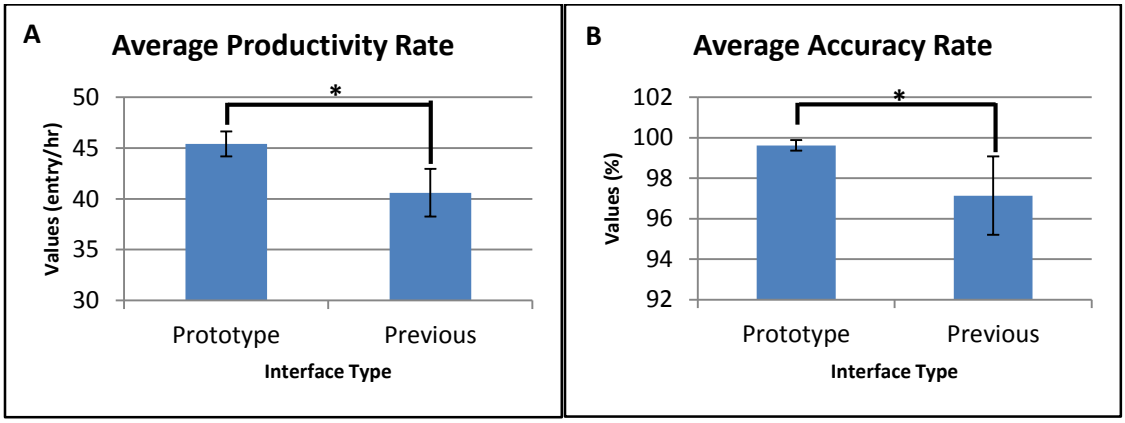


Figure 8. Semester statistics of undergraduate biocuraters. A. Average productivity rate per associate (mean ± SD) of new database interface prototype and previously used database interface (n=22, * $p<0.05$). B. Average accuracy rate per associate (mean ± SD) of new database interface prototype and previously used database interface (n=22, * $p<0.05$).

DISCUSSION

Undergraduate students have proven to be both efficient and accurate in their work in biocuration. With their high level of productivity and accuracy levels on par with Ph.D level employees, the use of undergraduates in the biocuration process could be further utilized as shown through the work of this study. When given enforced productivity and accuracy requirements and rigorous training, undergraduate students in this lab have been successful in extracting data from biomedical papers, adhering to the high standards set for them. The reconstruction of the FileMaker database has helped streamline the process, in turn increasing the speed and accuracy of the process, making it coherent and effective to extract useful data. While further research should be done on other labs employing similar techniques, the results of this study show that in combination with an effective database and proper training, undergraduate students are able to perform biocuration at the desired standards that are currently being performed by subject matter experts.

As shown through the database prototype analysis, productivity and accuracy are typically positively correlated, and high accuracy can be achieved even at higher levels of productivity. With an ergonomic user interface with proper training and strict quality control, higher quantity does not sacrifice the quality of the work provided. As undergraduates demonstrate the ability to complete biocuration at a level that is comparable to Ph.D level biocurators, these skilled professionals can spend more time analyzing the collected data rather than collecting it themselves. This will lead to a greater availability in making stride in well published ALS research areas.

This study focused on the training of undergraduates and their interface with a reconstructed G93A SOD1 transgenic ALS mouse database; however, results of this study are applicable for biocuration of other diseases. Not only can undergraduates perform biocuration with the proper database, they can perform complex biocuration methods while maintaining high standards of productivity and accuracy. With an accuracy ranging from 99-100%, adequate training and quality management of undergraduate biocurators using an ergonomic user interface has proven to be successful, opening the field to many future possibilities. As technology continues to improve, it is hoped that biocuration could eventually be a completely automated process, thus eliminating any human error in the process and fully optimizing rates of productivity.

ACKNOWLEDGMENTS AND CONTRIBUTIONS

I would like to extend a huge thank you to the following people in the Amyotrophic Lateral Sclerosis Lab for their contribution and support for this research and thesis report:

Mentor: Dr. Cassie S. Mitchell

Database Creation: Keval Tilva, T. Clark Howell, Renaid Kim, Rachel True

Biocuration Protocol: Ashlyn Cates, Rachel True

Training Protocol: Rachel True

Quality Management Protocol: Ashlyn Cates, Rachel True

Product Assessment: Rachel True

Biocuration: All associates of the ALS Lab

REFERENCES

Acsadi, G., Anguelov, R. A., Yang, H., Toth, G., Thomas, R., Jani, A., . . . Lewis, R. A. (2002). Increased survival and function of SOD1 mice after glial cell-derived neurotrophic factor gene therapy. *Human gene therapy, 13*(9), 1047-1059.

Bandrowski, A. E., Cachat, J., Li, Y., Muller, H. M., Sternberg, P. W., Ciccarese, P., . . . Martone, M. E. (2012). A hybrid human and machine resource curation pipeline for the Neuroscience Information Framework. *Database (Oxford), 2012*, bas005. doi: 10.1093/database/bas005

Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., Eppig, J. T., & Mouse Genome Database, G. (2011). The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. *Nucleic Acids Res, 39*(Database issue), D842-848. doi: 10.1093/nar/gkq1008

Burge, S., Attwood, T. K., Bateman, A., Berardini, T. Z., Cherry, M., O'Donovan, C., . . . Gaudet, P. (2012). Biocurators and biocuration: surveying the 21st century challenges. *Database (Oxford), 2012*, bar059. doi: 10.1093/database/bar059

Chiba, T., Yamada, M., Sasabe, J., Terashita, K., Aiso, S., Matsuoka, M., & Nishimoto, I. (2006). Colivelin prolongs survival of an ALS model mouse. *Biochemical and biophysical research communications, 343*(3), 793-798.

Drabkin, H. J., Blake, J. A., & Mouse Genome Informatics, D. (2012). Manual Gene Ontology annotation workflow at the Mouse Genome Informatics Database. *Database (Oxford), 2012*, bas045. doi: 10.1093/database/bas045

Elmasri, R., Navathe, S. B. (2010). *Fundamentals of Database Systems (6th Edition)* (6 ed.): Addison-Wesley.

Gilchrist, C., Gray, D., Stieber, A., Gonatas, N., & Kopito, R. (2005). Effect of ubiquitin expression on neuropathogenesis in a mouse model of familial amyotrophic lateral sclerosis. *Neuropathology and applied neurobiology, 31*(1), 20-33.

Guzzi, P. H., Veltri, P., & Cannataro, M. (2013). OntoPIN: an ontology-annotated PPI database. *Interdiscip Sci, 5*(3), 187-195. doi: 10.1007/s12539-013-0173-x

Hirschman, L., Burns, G. A., Krallinger, M., Arighi, C., Cohen, K. B., Valencia, A., . . . Winter, A. G. (2012). Text mining for the biocuration workflow. *Database (Oxford), 2012*, bas020. doi: 10.1093/database/bas020

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., . . . Rhee, S. Y. (2008). Big data: The future of biocuration. *Nature, 455*(7209), 47-50. doi: 10.1038/455047a

Lobo, D., Feldman, E. B., Shah, M., Malone, T. J., & Levin, M. (2014). Limbform: a functional ontology-based database of limb regeneration experiments. *Bioinformatics*. doi: 10.1093/bioinformatics/btu582

Lopez, L. D., Yu, J., Arighi, C. N., Huang, H., Shatkay, H., & Wu, C. (2011). *An automatic system for extracting figures and captions in biomedical pdf documents.* Paper presented at the Bioinformatics and Biomedicine (BIBM), 2011 IEEE International Conference on.

Lu, Z., & Hirschman, L. (2012). Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. *Database (Oxford), 2012*, bas043. doi: 10.1093/database/bas043

Mandler, R. N., Anderson, F. A., Jr., Miller, R. G., Clawson, L., Cudkowicz, M., Del Bene, M., & Group, A. C. A. R. E. S. (2001). The ALS Patient Care Database: insights into end-of-life care in ALS. *Amyotroph Lateral Scler Other Motor Neuron Disord, 2*(4), 203-208. doi: 10.1080/14660820152882214

Roussey, C., Pinet, F., Kang, M. A., & Corcho, O. (2011). An Introduction to Ontologies and Ontology Engineering. *Ontologies in Urban Development Projects*, 9-38. doi: Doi 10.1007/978-0-85729-724-2_2

Silani, V., Messina, S., Poletti, B., Morelli, C., Doretti, A., Ticozzi, N., & Maderna, L. (2011). The diagnosis of Amyotrophic lateral sclerosis in 2010. *Arch Ital Biol, 149*(1), 5-27. doi: 10.4449/aib.v149i1.1260

Smith, B., & Scheuermann, R. H. (2011). Ontologies for clinical and translational research: Introduction. *J Biomed Inform, 44*(1), 3-7. doi: 10.1016/j.jbi.2011.01.002

Taccioli, C., Tegner, J., Maselli, V., Gomez-Cabrero, D., Altobelli, G., Emmett, W., . . . Stupka, E. (2011). ParkDB: a Parkinson's disease gene expression database. *Database (Oxford), 2011*, bar007. doi: 10.1093/database/bar007

Takahashi, S., & Kulkarni, A. B. (2004). Mutant superoxide dismutase 1 causes motor neuron degeneration independent of cyclin-dependent kinase 5 activation by p35 or p25. *Journal of neurochemistry, 88*(5), 1295-1304.

Teng, Y. D., Benn, S. C., Kalkanis, S. N., Shefner, J. M., Onario, R. C., Cheng, B., . . . Snyder, E. Y. (2012). Multimodal actions of neural stem cells in a mouse model of ALS: a meta-analysis. *Sci Transl Med, 4*(165), 165ra164. doi: 10.1126/scitranslmed.3004579

Turner, B. J., Lopes, E. C., & Cheema, S. S. (2003). Neuromuscular accumulation of mutant superoxide dismutase 1 aggregates in a transgenic mouse model of familial amyotrophic lateral sclerosis. *Neuroscience letters, 350*(2), 132-136.

Wang, L.-J., Lu, Y.-Y., Muramatsu, S.-i., Ikeguchi, K., Fujimoto, K.-i., Okada, T., . . . Kume, A. (2002). Neuroprotective effects of glial cell line-derived neurotrophic factor mediated by an adeno-associated virus vector in a transgenic animal model of amyotrophic lateral sclerosis. *The Journal of neuroscience, 22*(16), 6920-6928.

Wu, A. S., Kiaei, M., Aguirre, N., Crow, J. P., Calingasan, N. Y., Browne, S. E., & Beal, M. F. (2003). Iron porphyrin treatment extends survival in a transgenic animal model of amyotrophic lateral sclerosis. *Journal of neurochemistry, 85*(1), 142-150.

Zhang, Z., Zhu, W., & Luo, J. (2014). Bringing biocuration to China. *Genomics Proteomics Bioinformatics, 12*(4), 153-155. doi: 10.1016/j.gpb.2014.07.001

APPENDIX 1 – ALS LAB TRAINING PROTOCOL

LEVEL 1:

1. **Paper Capture**

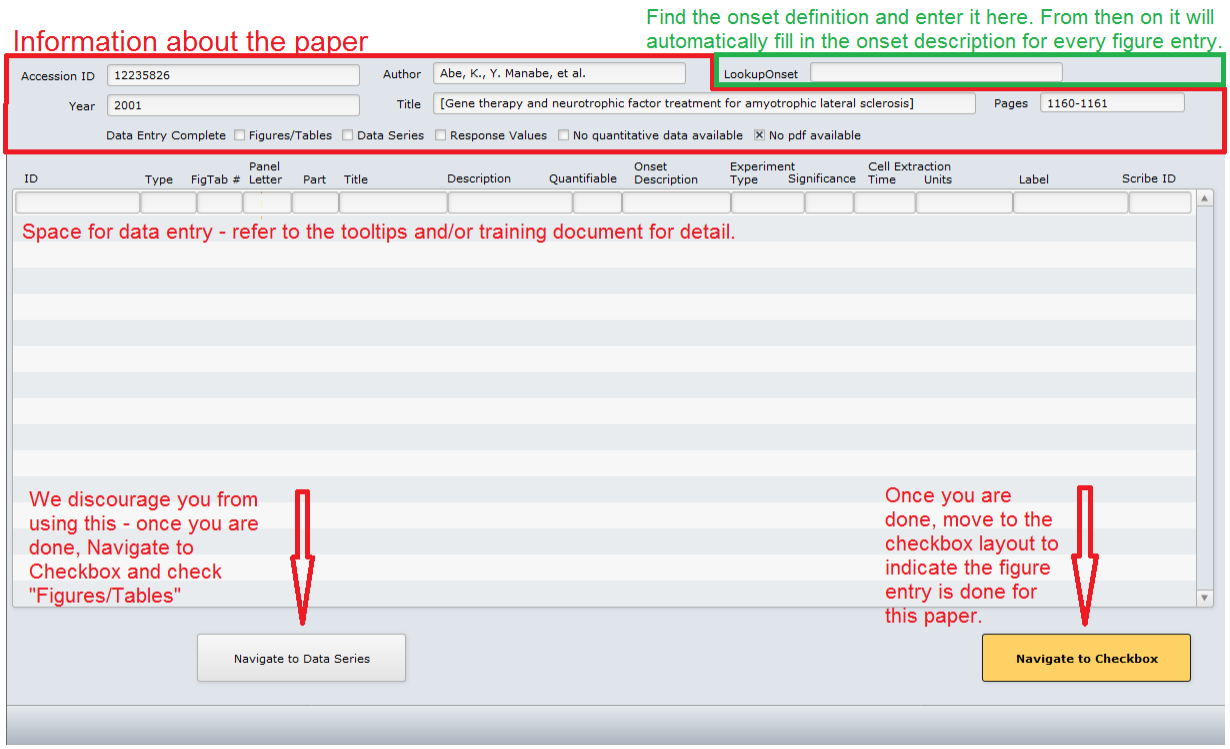
Step-by-step procedure

1. Open FileMaker
    a. Cancel out of 'Open New or Existing File' (if it appears)
    b. Click File>Open Recent>**ALS G93A Mouse (CSM-F-2013)**
2. Go to 'Papers' layout found in the drop down bar in the top left
3. Scroll through papers until a blank record is found.
    a. This means there are no boxes checked under 'Data Entry Complete'
4. Copy Accession ID
5. Go to '1-Paper Entry' layout
6. Type Accession ID into search bar. Ensure the correct paper is pulled up.
7. Open an Internet browser
    a. Access biomedical engineering databases (use Ga Tech ID or Emory ID)
    b. Find paper in one of the databases. Download the full PDF and save in the format 'Page Author Year.pdf' to the G93A Papers folder on the server
        i. Double Check this info matches what's written in FileMaker
    c. If there is no downloadable pdf, check the no pdf available box in FileMaker and add your scribe ID and move onto the next paper
8. Click the 'Navigate to Checkboxes' button. Check the 'Document Capture' box
9. **Double check your work**
10. Return to 'Papers' layout and find another paper
11. When ready to leave, make sure you close the programs and log off.

LEVEL 2:

1. **Figure/Table Capture:**

It is important than you refer to the flowchart for the general workflow, since not all figure panels are entered in. Some information may already have been filled (imported from the old database). For such quantifiable figures, you need to fill out the empty fields where applicable and write down your scribe ID.



ID: This is automatically generated. Do not type anything

Type: Select either Figure or Table

Part: Enter the part number if the figure or table has multiple parts. If all parts are non-quantifiable, you do not need to enter each part – this is okay to leave blank if necessary

Panel: Enter the panel letter if the figure or table has multiple panels. If all panels are non-quantifiable, you do not need to enter each panel – this is okay to leave blank if necessary

Title: Enter the title of the figure or table. Often the first sentence of the figure caption. The title is usually the same across the panels belonging to the same figure.

Description: Enter the description of the figure (usually follows the title in the caption. Each part/panel will likely have a different description)

Quantifiable: Select yes or no – if the figure or table has quantifiable (numeric) data you will select yes. If the figure does not have quantifiable data, you do not need to continue to enter the onset description, experiment type, e.g.

Onset Description: Generally there is one onset description for the entire paper, but sometimes they will change between figures. You can find this by scanning the paper. Examples are: Grip Failure, Rotarod

Failure, e.g. – this is okay to leave blank if necessary. If the onset description is the same for the entire paper, you may type it into the "LookupOnset" entry in the top right corner by "Author." This will automatically put it in every entry for the paper.

Experiment Type: Select Invivo, mixed, or Invitro. Invivo means "within the living" and Invitro means "in glass". If an experiment uses a whole living mouse, this would be invivo. If an experiment uses a cell culture from a live mouse at different timepoints this could be classified as mixed. If there is a cell extraction time specified, it is most likely mixed. If an experiment takes one extraction from a dead or live mouse or uses components of an organism that have been isolated from their usual biological surroundings, this would be invitro. If the timespan is a few hours it is likely to be invitro. Invitro has to start with a cell culture.

Significance: Select yes or no – if the figure has any statistical significance. Usually indicated by an asterisk on the graph.

Cell Extraction Time: If the experiment is mixed, there may be a "time point" at which the cells were extracted, we want to capture this data here and not under timepoint. For example: If an experiment extracts cells from a mouse at 90 days and then does an experiment and monitors the results every 5 days, the cell extraction time would be 90 days and the timepoints would be 90 days, 95 days, 100 days, etc. If this is not specified in the paper, the experiment is likely to be just invitro rather than "mixed."

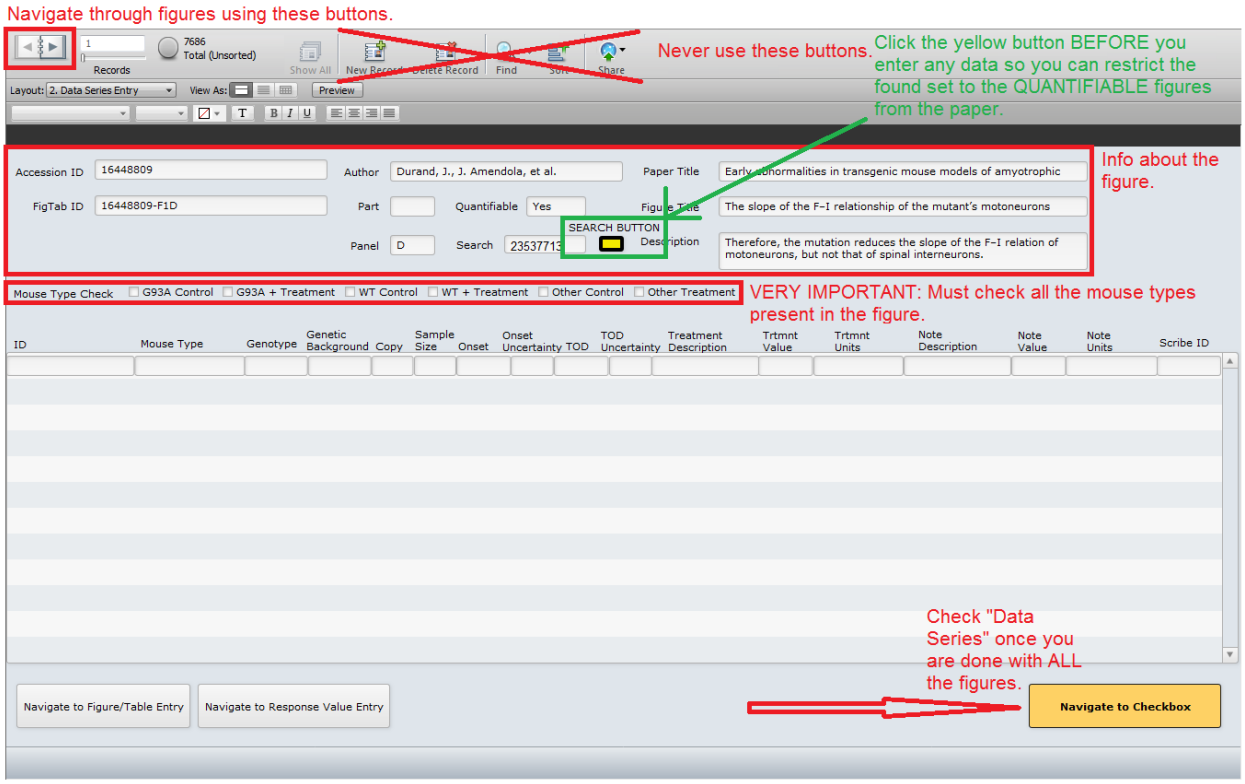
Cell Extraction Units: Select the units for the cell extraction time – days, weeks, etc.

Label: Select "No Wild Type" or "All Four Types" if the figure has either of these cases. It is okay to leave this field blank if necessary.

Scribe ID: Select your scribe ID

LEVEL 2:

1. **Data Series Entry:**



Mouse Type Check: Check the types of mice present in the figure. If there are "Other" types select the other mouse checkbox and enter the mouse type in the data series note. (Example: G85R, G37R, G85A, G41S, etc.).

Mouse Type: Select the mouse type. Before leaving the Data Series Entry, make sure you have an entry for every mouse type check box you checked. If you checked "Other" and that type has not be previously been documented, click Edit in Mouse Type and type it in following the same format as other entries.

Genotype: Select the genotype – hemizygous, homozygous, or heterozygous. You'll find this in the paper and probably not in the figure description. Most of the time the genotype is the same throughout the paper. – this is okay to leave blank if necessary

Current conventions:

Homozygous: methods list only founder strain (the long-name strain designations from the Jackson pages) with no reference to cross-breeding with other mouse types.

Hemizygous: Founder strains backcrossed on wild type mice (Non G93A C57BL/6, B6SJL, and others). Very commonly used and is only sometimes explicitly stated.

Heterozygous: Explicitly stated as heterozygous or a result of crossing with other transgenic strains (AKA "double transgenics").

Genetic Background: Select either B6SJL or C57BL/6 – this is usually listed with the laboratory (most mice come from Jackson Laboratory so if you search for "Jackson" you might be able to find this). You'll

find this in paper and probably not in the figure description. – Leave it blank if you can't find it. Genetic background refers to the line used for hemizygous breeding patterns in the control SOD1 G93A mice used in the paper. The "B6SJL" line has a reported 50% survival of 128.9 +/- 9.1 days, see the Jackson mouse page for more info (http://jaxmice.jax.org/strain/002726.html). The C57BL/6 background mice are sometimes designated as "B6.Cg-Tg(SOD1-G93A)1Gur/J" and are sometimes referred to as "B6 mice" within the paper. The designation is only specified directly in the breeding section of the methods. BE CAREFUL. The 50% survival for these mice is 157.1 +/- 9.3 days as per the Jackson page (http://jaxmice.jax.org/strain/004435.html). Both mouse types are proper G93A mice and are both capable of high/low copy variations.

Copy: Either high or low copy genes. B6SJL-Tg low copy mice become paralyzed in one or more limbs beginning around 183-210 days of age. Life expectancy is normally four to six weeks beyond onset of symptoms. High copy mice die sooner; High-copy B6SJL-Tg mice have an abbreviated life span: 50% survive at 128.9+/-9.1 days (in contrast to high-copy C57BL/6J background where 50% survival is observed at 157.1+/-9.3 days). Only report as low copy number if it is specifically stated in the paper or the Jackson page for the strain details the high/low copy number. High copy is most common and is "default". Low copy will be specifically stated, but leave blank if unable to determine. High copy mice are occasionally referred to as "G1H" regardless of genetic background. Example low copy strain designations: B6SJL-Tg(SOD1*G93A)^{dl}1Gur/J (http://jaxmice.jax.org/strain/002300.html) and B6.Cg-Tg(SOD1*G93A)^{dl}1Gur/J (http://jaxmice.jax.org/strain/002299.html) , the designation "dl" specifies low copy number .

Mouse Count: Enter the number of mice used in the experiment. Typically denoted as n=# -- leave blank if unfound.

Onset: Disease onset time in days. If you need to convert, remember there are 52 weeks in a year. Please do not convert it using months.

Onset Uncertainty: Disease onset uncertainty **in days**. For example if the onset is listed as 120+/-3 days, you would list 3 as the uncertainty. Do NOT include the +/- sign.

Time of Death (TOD): Time of Death **in days**. Again, convert if necessary.

TOD Uncertainty: Time of Death uncertainty **in days**. Do not include the +/- sign.

Treatment Description: Enter the treatment description. The treatment is the independent variable that is foreign and causes a response. For clarification this means if there is an "Ab Titer" used which causes a response this would be listed as the treatment description – this can be left blank if necessary

Treatment Value: The value of the treatment – this can be left blank if necessary

Treament Units: The units that describe the treatment value – this can be left blank if necessary

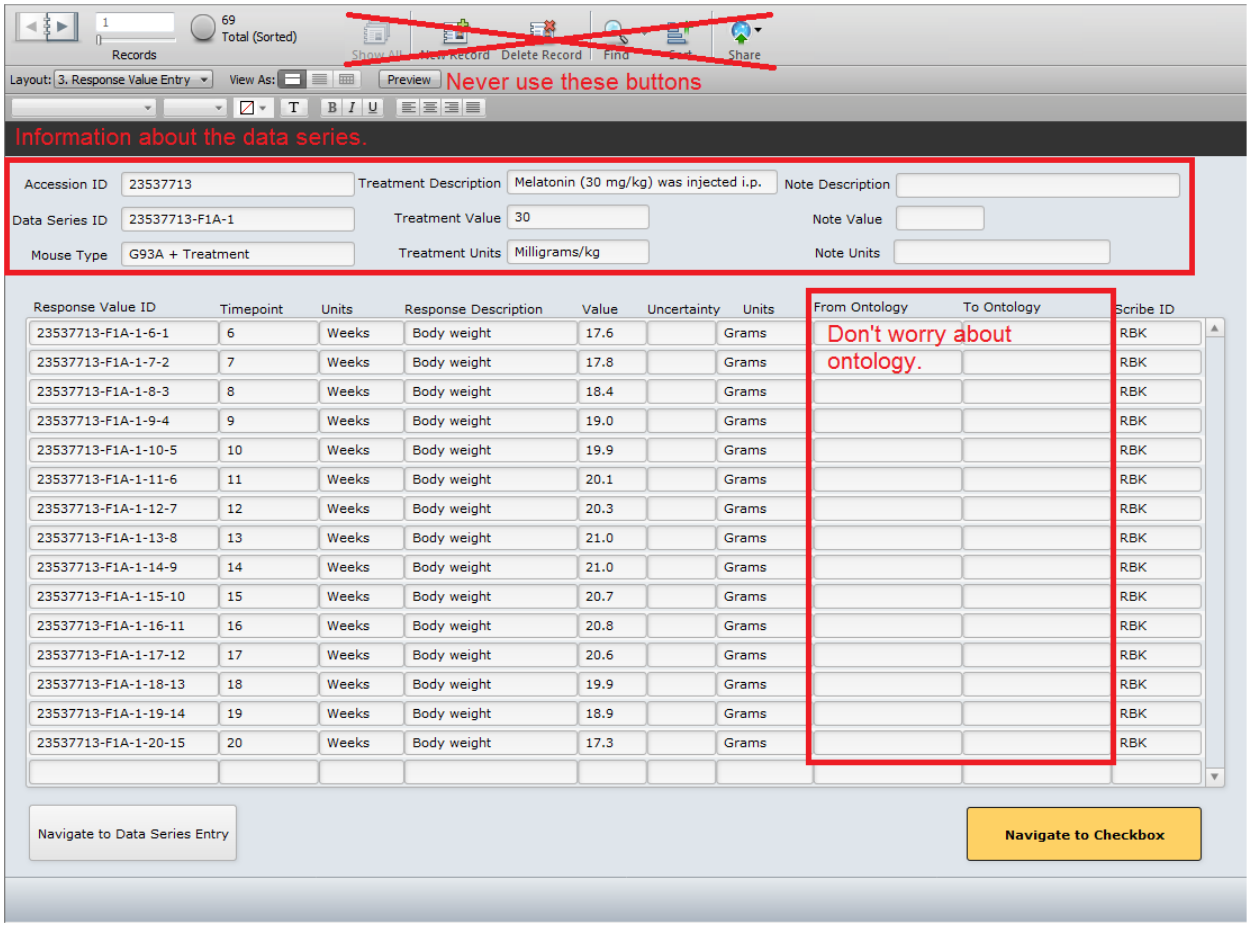
Note Description: Enter the note description. The note another variable or independent variable that may not cause the change, but can describe what is changing between data points. For clarification this means if the experiment is measuring the GFAP cell density in relation to the distance from a cluster of nerve cells, the distance would be the note and the GFAP cell density would be the response. An experiment can have both a treatment and a note. If the gender of the mouse is specified, please enter this in the note – this can be left blank if necessary

Note Value: Enter the note value that pertains to the note – this can be left blank if necessary

Note Units: Enter the units that describe the note value – this can be left blank if necessary

Scribe ID: Select your Scribe ID

2. Response Value Entry:



Timepoint: The time at which data were taken.
Units: Select the units that describe the timepoint
Response Description: This previously was the "measured" field. This is what is measured for the response. This can be frequency, count, weight, e.g. If it's a graph, response description is most likely the Y-axis.
Value: The value related to the response description that describes the value of the response in relation to the timepoint
Units: Enter the units that describe the value
From: **LEAVE BLANK**. It is reserved for gain analysis.
To: **LEAVE BLANK**. It is reserved for gain analysis.
Scribe ID: Select your Scribe ID

3. Checkbox Entry

Make sure to check the appropriate checkbox after each stage of the entry.