Manuscript targeting publication in F1000 Research

KnowMore: An Automated Knowledge Discovery Tool Developed During the 2021 NIH SPARC Codeathon

Ryan Quey[1], Matthew A. Schiefer[2, 3, 4], Anmol Kiran[5,6], Bhavesh Patel[7,*]

Affiliations

1 Anant Corporation, Washington, D.C., USA

2 Malcom Randall VA Medical Center, Gainesville, FL, USA

3 Department of Biomedical Engineering, University of Florida, Gainesville, FL, USA

4 SimNeurix, LLC, Gainesville, FL, USA

5 Malawi-Liverpool-Wellcome Trust, Blantyre-3, Malawi

6 Institute of Infection, Veterinary & Ecological Sciences, University of Liverpool, Liverpool CH64 7TE, UK

7 California Medical Innovations Institute, San Diego, CA, USA

*Email: bpatel@calmi2.org

Abstract

This manuscript provides methods and outcomes of KnowMore, the automated knowledge discovery tool developed by our team during the 2021 SPARC Codeathon. Currently, the process of comparing and analyzing multiple SPARC datasets is tedious as it requires manually investigating each dataset of interest and then requires downloading them before being able to do any cross-analysis. It is crucial to enhance this process to enable rapid discoveries across SPARC datasets and encourage more researchers to use the SPARC Data Portal. To fill this need, we propose KnowMore, an automated knowledge discovery tool integrated into the SPARC Portal that only requires the user to select their datasets of interest and then hit a "Discover" button to launch the automated discovery process. KnowMore then makes use of several SPARC resources in the back-end (Pennsieve, osparc, SciScrunch, protocols.io, Biolucida) to generate various visualization items that are intended to help the user identify potential similarities, differences, and relations across the datasets and their data files, which could lead to a new discovery, new hypothesis, or simply guide the user to the next logical step in their discovery process. The outcome of this project is a code architecture ready to onboard the SPARC Portal and help researchers desiring to reuse SPARC datasets. The program has been built and documented such that more and more visualization items could be easily added. The potential for automated discoveries from SPARC datasets is huge given the unique SPARC data ecosystem, and KnowMore has only demonstrated a small highlight of what could be achieved at a much larger scale to speed up discoveries.

Keywords

Metadata, FAIR, Natural Language Processing, Knowledge graph, Vue, Python

Introduction

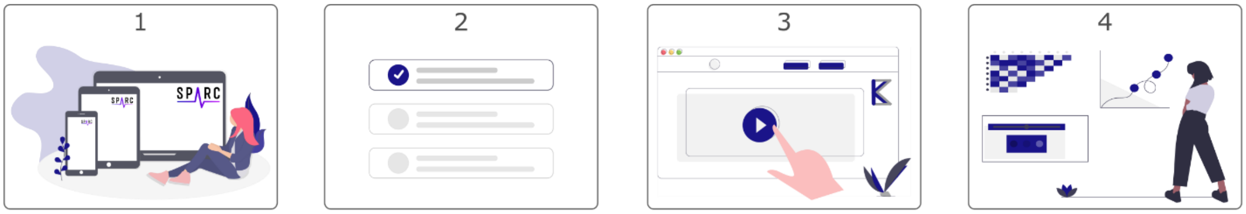
The NIH's Stimulating Peripheral Activity to Relieve Conditions (SPARC) program seeks to accelerate the development of therapeutic devices that modulate electrical activity in nerves to improve organ function. A major focus of the SPARC program is the generation of rich datasets that provide key resources for understanding nerve-organ interaction and guiding the development of neuromodulation therapies. These datasets are publicly available through an open data platform, the SPARC Portal (sparc.science). As of July 2021, 115 datasets are available spanning multiple scales (cellular, tissue, organ level), organs (stomach, large intestine, small intestine, heart, bladder, urinary tract, lung, pancreas, spleen), species (pig, human, rat, mouse, dog), data types (scaffold data, histology, immunohistochemistry, electrical impedance tomography, 3D microscopy, morphometric analyses, computer simulations of single axons or populations of axons, electrophysiological responses to electrical stimulation).

To ensure SPARC datasets are Findable, Accessible, Interoperable, and Reusable (FAIR), they are curated according to the SPARC Data Structure (SDS)[i], which uses FAIR Data standards[ii] designed to capture the large variety of data generated by SPARC investigators. The SPARC program thus provides a wealth of openly available and well-curated datasets via the SPARC Portal.

The sparc.science portal provides several means of accessing data. A standard search feature is available. Alternatively, the user can find datasets by browsing through data categorized by organ system. The user can also use an interactive map to click on organs or nerves of interest in animal models of interest and the website will provide links to associated datasets. These pathways make it easy to find data. Clicking on a link provides the user with details about the study and links to get all or portions of the dataset.

While it is very easy to look at the details of any single SPARC dataset on the portal, there is currently **no easy way to make rapid analysis across multiple datasets of choice**, leaving the user to open multiple windows and tile them across the screen to visually compare images Typically, a researcher wanting to find relations across datasets would have to do so manually, i.e., read the description of each dataset, go through each protocol, browse files that are accessible from the browser, etc. If they find that the datasets could have some interesting relations that is worth investigating further, they will then have to download each dataset (payment may be required for large datasets according to AWS pricing) before analyzing them further. Depending on the format of the data, this may require programming skills beyond that of many users. After spending time collating data in a form that allows comparison across the different datasets, the user may find that, in fact, the data sets did not contain the information they needed. Currently, the process of comparing datasets is tedious and needs to be urgently improved to 1) enable rapid discoveries across SPARC datasets and 2) encourage more researchers to use the SPARC Data Portal.

To address this problem, we have developed a tool called KnowMore. KnowMore is an automated knowledge discovery tool integrated within the SPARC Portal. With just a few clicks, KnowMoreallows users of the portal to visualize potential relations, similarities, correlations, and differences between several key features of multiple SPARC datasets. KnowMore performs text mining, provides a summary table, and plots data that is common across all selected datasets, thus providing the user with a quick means of determining if all or some of the datasets are appropriate for their research. This simple process is illustrated in Figure 1.



Go to sparc.science from any web connected devices

Select SPARC datasets that you want to analyze together

Launch the discovery process from the KnowMore tab with a single click

Visualize potential similarities, differences, and connections between the datasets

Figure 1. Usage of KnowMore

This manuscript describes the structure of KnowMore and provides an example of insight provided by the tool when applied to a set of three sample datasets that constituted our use case for demonstration during the 2021 SPARC Codeathon.

Methods

Architecture

The overall workflow of KnowMore is shown in the Figure 2. Our architecture consists of three main blocks that can all run independently:

1. The front end of our app is based on a fork of the sparc-app (i.e. the front-end of sparc.science) where we have integrated additional UI elements and back-end logic for KnowMore. Learn more about the sparc app.
2. The back-end consists of a Flask application that listens to front-end requests and launches the data processing jobs.

3. The data processing and result generation is done through a Matlab code (for 'MAT' data files) and a Python code (all other data types) that run on osparc, the SPARC supported cloud computing platform. Learn more about osparc.

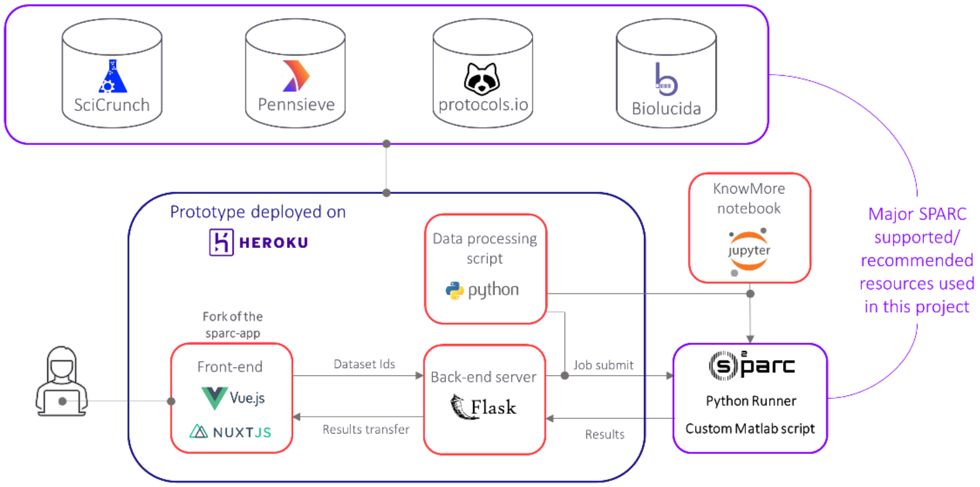


Figure 2. Workflow

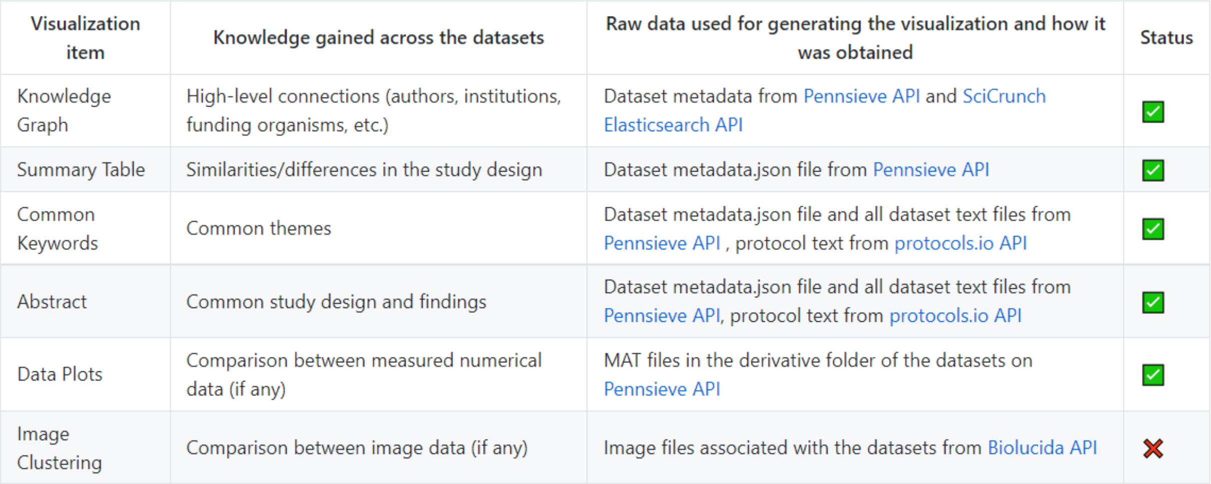
Such a design was motivated by our aim of making KnowMore ready to on-board the SPARC Data Portal:

- Integrating the front-end of KnowMore would only require merging our fork of the sparc-app with the main branch sparc-app branch.
- The back-end of the sparc-app, the sparc-api, is build with Flask so the KnowMore back-end would be readily compatible.
- The data processing jobs are designed to run on osparc, the SPARC supported cloud computing platform, and would not require any type of integration as our back-end ensures communication with ospar

Moreover, each of the three main elements of KnowMore is fully independent. While the front-end will not be of much use on its own, having the back-end fully interoperable is very valuable as our flask application can be connected to any front-end if needed (another analysis tool, website, software, etc.). The data processing and results generation jobs are also independent such that they can be used directly to get the visualization items. We have demonstrated that by developing a Jupyter Notebook that communicates directly with osparc to run the knowledge discovery jobs based on user-specified dataset Ids. Note that the data for the Knowledge graph is obtained from Pennsieve/Scicrunch on the front-end for efficiency but the same results can be generated in the back-end as well.

Outputs and Data Processing

The output of KnowMore consists of multiple interactive visualization items displayed to the user such that they can progressively gain knowledge on the potential similarities, differences, and relations across the datasets. This output is intended to provide foundational information to the user such that they can rapidly make novel discoveries from SPARC datasets, generate new hypotheses, or simply decide on their next step (assess each dataset individually on the portal, download and analyze the datasets further, remove/add datasets to their analysis pool, etc.). A list of the visualization items is provided in the table below, along with the potential knowledge that could be gained from each of them.



Visualization item	Knowledge gained across the datasets	Raw data used for generating the visualization and how it was obtained	Status
Knowledge Graph	High-level connections (authors, institutions, funding organisms, etc.)	Dataset metadata from Pennsieve API and SciCrunch Elasticsearch API	☑
Summary Table	Similarities/differences in the study design	Dataset metadata.json file from Pennsieve API	☑
Common Keywords	Common themes	Dataset metadata.json file and all dataset text files from Pennsieve API , protocol text from protocols.io API	☑
Abstract	Common study design and findings	Dataset metadata.json file and all dataset text files from Pennsieve API, protocol text from protocols.io API	☑
Data Plots	Comparison between measured numerical data (if any)	MAT files in the derivative folder of the datasets on Pennsieve API	☑
Image Clustering	Comparison between image data (if any)	Image files associated with the datasets from Biolucida API	✗

The process of getting these outputs starts simply with the Ids of the Datasets selected by the user which are obtained using the Pennsieve API. From there, we leverage several SPARC-supported and recommended resources to collect the raw data required to generate the above-mentioned outputs. Details about each of the visualization items are provided below.

Knowledge Graph

The knowledge graph is built with metadata obtained for each dataset using the SciCrunch Elasticsearch API. The visualization library Vega is used to display the graph interactively.

Summary Table

The summary table is built with information collected from the metadata.json file of each dataset, which is a standard file generated for each SPARC datasets when published. The file was streamed into our program using the Pennsieve API. The visualization library Plotly is used to display the table interactively.

Keywords

Text for identifying common keywords from the datasets is obtained, for each dataset, from different sources:

- The description included in metadata.json file using the Pennsieve API
- Text from all the text files in the dataset using the Pennsieve API,
- Text from the protocol on protocols.io associated with the dataset using the protocols.io API. The link to the protocol.io protocol is collected from the metadata.json file of the dataset

The text is then combined together to create a large paragraph for each dataset. Natural Language Processing python library NLTK is then used to clean the text (e.g., remove stopwords (is, an, the etc)). Biological keywords words were identified using spaCy python module and scispacy model. ScispaCy contains models related to using spaCy for scientific documents. The frequency of biological words was counted for each dataset. The final frequency of the keywords was assigned based lowest occurrence among the datasets. Twenty most frequent words were selected from the generated word frequency table.

Abstract

Paragraphs generated from datasets for the keywords identification were merged and divided into sentences. Each sentence was further divided into words and stopwords were removed. Frequency of each remaining word in a sentence was calculated by counting and converted in to vectors where keywords represented the direction and frequencies represented magnitude. The distance of two sentences was calculated using equation (1 – cosine similarity). Where cosine similarity is expressed as

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Where A and B are words frequency in vectors of two sentences. Based of pairwise distance of sentence pagerank was assigned to each sentence using python networkX module and sentences were ordered based in pagerank in decrease order. For the abstract, top 10 highest ranked sentences are selected.

Data Plots

For datasets with data saved in .mat format, the MATLAB script main.m will collate the data into a data table. The script next determines which columns in the data table can be used for plotting purposes. Columns containing categorical data are limited to the x-axis. Columns containing numerical data that can be plotted on either the x or y axis. Columns containing any other type of data are excluded. Plots are then created of every variable that can be plotted on a y-axis against every variable that can be plotted on an x-axis. The MATLAB script outputs an Excel file that lists each of the plots created and the variables included in each plot. The Excel file also includes data for each plot. Additionally, MATLAB creates a json file that includes all data for each plot.

Resource Availability

Use Case

Setup

KnowMore was developed and tested mainly using three datasets available at sparc.science (Table 1). These datasets were selected because they have a common theme – quantified vagus nerve morphology – and span three species: rat, pig, and human. In principle, KnowMore is not specifically designed around these datasets and is coded to work with any user-selected datasets. However, for demonstration purposes, the data plots are currently limited to only appear when working with one or more of the three datasets listed in Table 1. Reasons for this are addressed in the Challenges section below and recommendations are put forth to expand the usability of this feature and increase the interoperability of SPARC datasets.

Table 1: Datasets used

ID	Title
60	Quantified Morphology of the Rat Vagus Nerve[iii]
64	Quantified Morphology of the Pig Vagus Nerve[iv]
65	Quantified Morphology of the Human Vagus Nerve with Anti-Claudin-1[v]

Initiating a KnowMore analysis requires five steps:

1. Use the search feature or browse for possible datasets of interest at sparc.science.
2. As datasets are identified that the user wants to compare, click on the "Add to KnowMore" button, visible in the header of the datasets. This will add the datasets to the KnowMore analysis.
3. Go to the KnowMore tab at the top of the webpage and check that all of the desired datasets are listed.
4. Decide which output to display. All possible output is displayed by default.
5. Click on the "Discover" button to initiate the automated analysis.

The number of datasets selected will affect the duration of time it takes to run the full discovery analysis. For the three datasets listed above, the time to display results in each section are listed in Table 2.

Table 2: Time required to display KnowMore results for the three datasets listed in Table 1.

KnowMore Section	Typical Time to Display
Knowledge Graph	<5 s
Summary Table	2.5 min
Common Keywords	2.5 min
Abstract	2.5 min
Data Plots	4 min
Image Clustering	Not Yet Available

Output

Knowledge Graph

The Knowledge Graph provides an interactive tool to visualize metadata across the three datasets (Figure 4). This provided the ability to quickly determine, for example, that all three datasets had four investigators in common (Cariello, Grill, Goldhagen, and Pelot) affiliated with Department of Biomedical Engineering at Duke and that the human dataset had additional investigators (Ezzell and Clissold) affiliated with the Department of Cell Biology and Physiology at the University of North Carolina. Hovering over any investigator's name provided that investigator's ORCID information. Hovering over any dataset provided its DOI information.

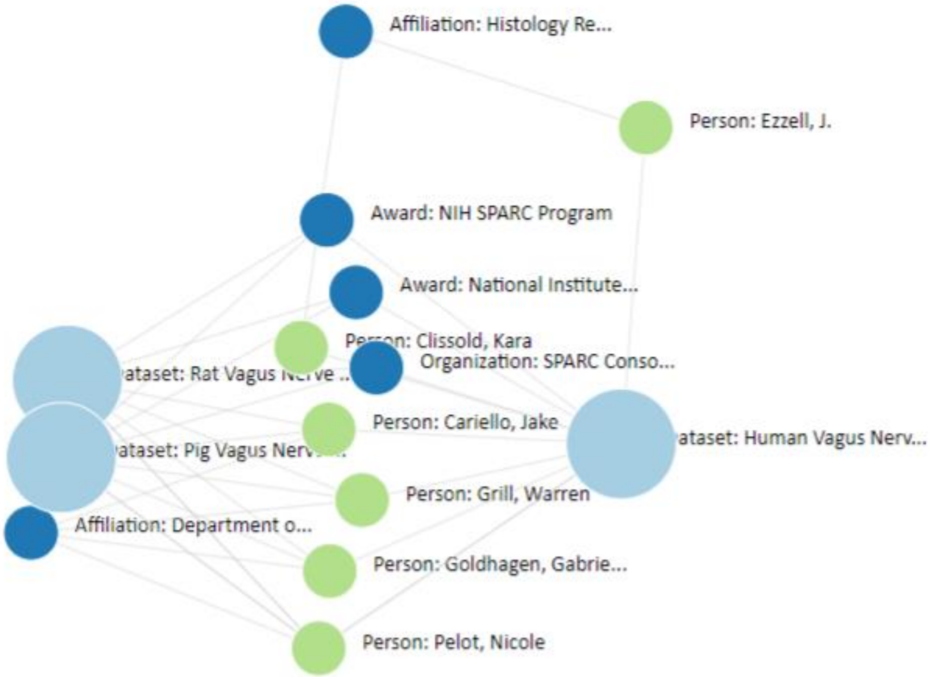

Affiliation: Histology Re...
Person: Ezzell, J.
Award: NIH SPARC Program
Award: National Institute...
Person: Clissold, Kara
Organization: SPARC Conso...
ataset: Rat Vagus Nerve ...
Person: Cariello, Jake
ataset: Human Vagus Nerv...
ataset: Pig Vagus Nerv...
Person: Grill, Warren
Affiliation: Department o...
Person: Goldhagen, Gabrie...
Person: Pelot, Nicole

Fig 4: KnowMore Knowledge Graph output for the three datasets listed in Table 1.

Summary Table

The Summary Table provides the user with key pieces of information from each study in tabular format (Table 3). From this table, the user can easily determine that datasets have several common metrics. However, perineurial thickness is not quantified in dataset 64.

Table 3: KnowMore Summary Table output for the three datasets listed in Table 1

Dataset ID	60	64	65
Title	Quantified Morphology of the Rat Vagus Nerve	Quantified Morphology of the Pig Vagus Nerve	Quantified Morphology of the Human Vagus Nerve with Anti-Claudin-1
Subtitle	Binary traces from segmentation of cross sections of cervical and subdiaphragmatic rat vagus nerves stained with Masson's trichrome. Quantified effective nerve diameter, effective fascicle diameter, number	Binary traces from segmentation of cross sections of cervical and subdiaphragmatic pig vagus nerves stained with Masson's trichrome. Quantified effective nerve diameter, effective	Immunohistochemistry micrographs of human vagus nerves labeled with anti-claudin-1. Binary traces from segmentation to quantify effective nerve diameter, effective fascicle diameter, number

	of fascicles, and perineurium thickness.	fascicle diameter, and number of fascicles.	of fascicles, and perineurium thickness.
Publication Date	2020-09-30	2020-10-01	2020-10-01
Number of Subjects	10	11	15
Species	Rattus norvegicus	Sus scrofa domesticus	Homo sapiens
Age	75 days - 268 days	10.5 weeks - 15 weeks	54 years - 90+ years
Sex	Female, Male	Female, Male	Female, Male
Number of samples	18	18	20
Specimen Type	vagus nerve	vagus nerve	vagus nerve
Anatomical Location(s)	left cervical vagus nerve; 11 mm from carotid bifurcation, subdiaphragmatic vagus nerve; 8.5 mm from esophageal hiatus and 8.5 mm from gastroesophageal junction; hepatic branch 10 mm from esophageal hiatus.	left cervical vagus nerve; 15 cm from bottom of jaw to top of sternum; sample middle ~2 cm; 6 cm from middle of sample to carotid bifurcatoin, left cervical vagus nerve; 13 cm from bottom of jaw to top of sternum; sample middle ~2 cm; 5 cm from middle of sample to carotid bifurcation.	left cervical vagus nerve; 35 mm from carotid bifurcation, left cervical vagus nerve; 20 mm from carotid bifurcation.

Common Keywords

The Common Keywords figure provides a graphical depiction of words that show up multiple times across the selected datasets (Figure 5). This size of the word in the image provides a visual representation of the weight (or frequency) of that word across the datasets. Not surprising, "nerve" is a large word as it shows up many times.

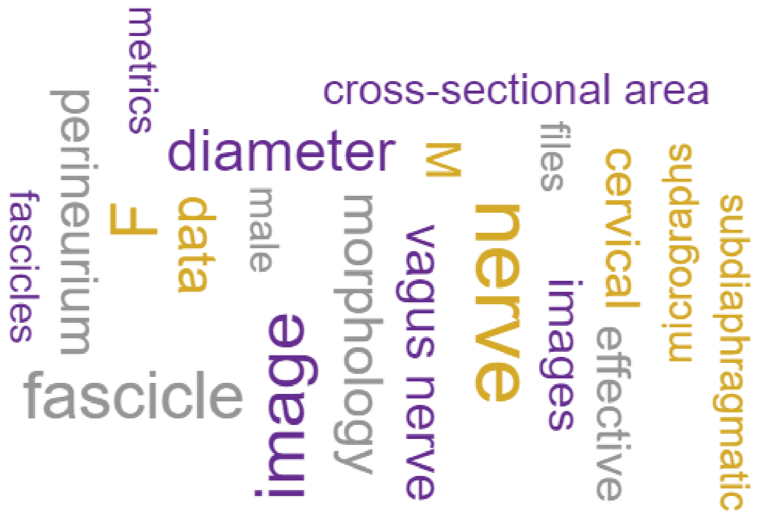


Fig 5: KnowMore Common Keywords output for the three datasets listed in Table 1.

Abstract

KnowMore generates a combined abstract that provides an overview of all datasets included in the study. The text is more detailed than the prior three outputs and can easily be copied and pasted into other documents. Additionally, KnowMore generates a heatmap illustrating the correlation between the studies based on the words used in the descriptions of these studies (Figure 6). This figure can guide the user in selecting highly correlated studies or eliminating studies that do not correlate well.

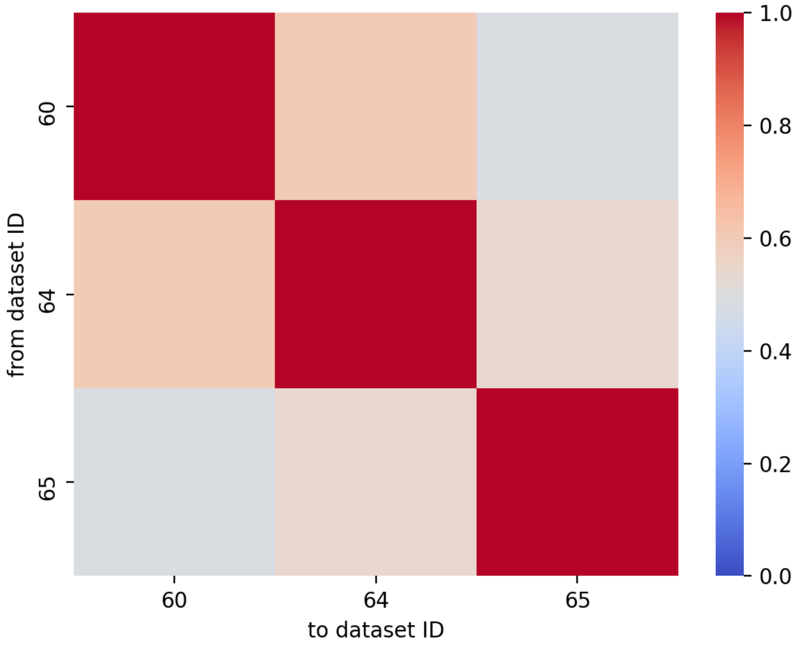


Fig 6: Correlation of the words used to describe the three datasets listed in Table 1.

Data Plots

For this use case, KnowMore also generates 20 scatter plots. Data points are color-coded to each dataset. Each axis is labeled with the variable being plotted. The variable name is obtained directly from the datasets. Three of the plots are presented here (Figure 7). Plot 3.4 reveals that pigs contain far more fascicles in their vagus nerves than do humans and humans contain more fascicles than rats. Plot 3.4 also reveals that pigs and rats have similar variability (spread) in their fascicle diameters whereas humans have a much greater spread in their fascicle diameters. Finally, Plot 3.4 illustrates that humans can have much larger fascicles than pigs. Plot 3.5 reveals that humans and pigs have similar-sized nerves, though pigs may, on average, have larger nerves. Plot 3.5 also reveals that the number of fascicles in the nerve tends to be greater for nerves of larger diameter within each species. That is, there appears to be a positive correlation between the number of fascicles in the nerve and the diameter of the nerve. However, Plot 4.5 suggests that there may not be a trend between the fascicle diameter and the nerve diameter.

Although these findings have been previously reported in some form[vi], the Data Plots can become a very useful tool in helping researchers quickly understand the underlying data across multiple datasets.

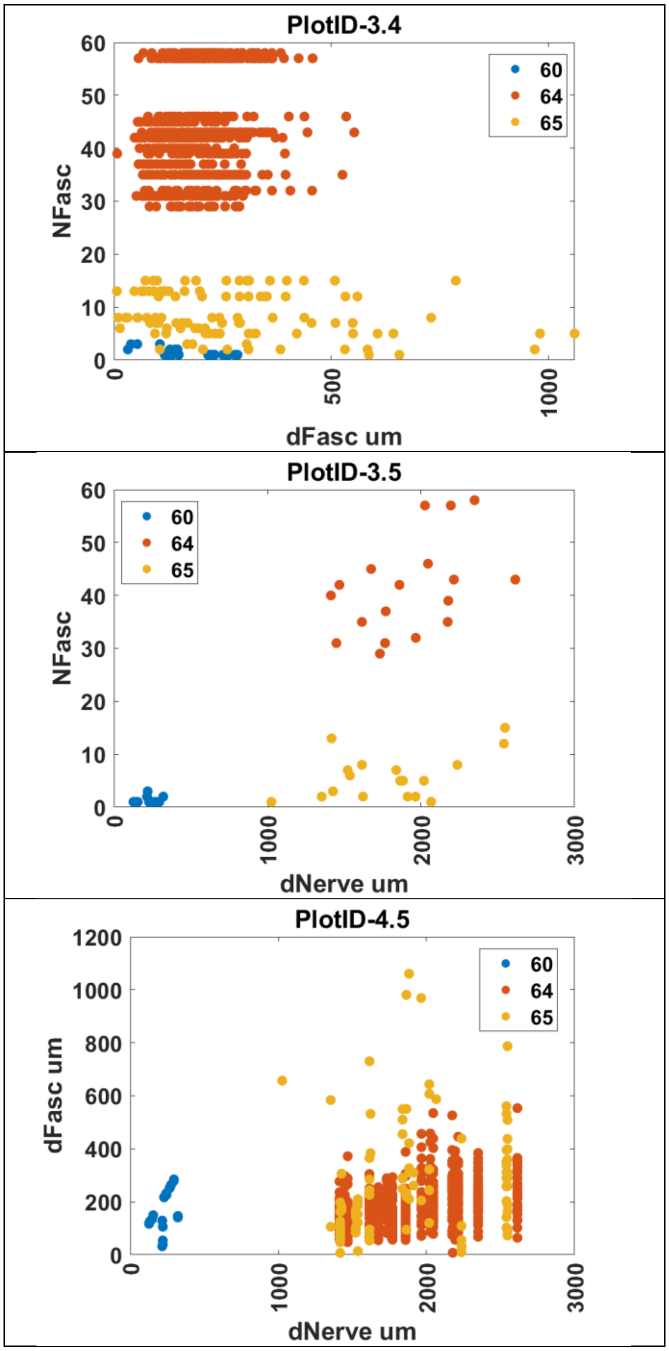


Fig 7: Three selected KnowMore Data Plots created from the three datasets listed in Table 1.

Conclusions and Next Steps

Potential for this tool

In less than 5 minutes, KnowMore can provide both a high-level metanalysis and a granular comparison across two or more studies on the SPARC portal. KnowMore outputs results at several levels depending on the needs of researcher. One can quickly determine personnel, institutional, and funding relationships between datasets. One can easily compare an overview of subjects included in the datasets and the techniques used to obtain data. Finally, if data are available for plotting, plots can reveal relationships within and across the studies that may reveal larger trends or help the researcher choose or eliminate particular datasets for more detailed analysis.

Challenges

SPARC has done an excellent job of standardizing the metadata associated with a study, and, as such, most of the KnowMore output is available across any selected studies. However, SPARC has not enforced data standardization. As such, the Data Plot output of KnowMore is currently limited to datasets that contain identical variable names stored in identical format. This is an uncommon occurrence across datasets.

Data can currently be stored in any number of formats. KnowMore's Data Plot currently requires data to be stored in a MATLAB .mat file, but this could be expanded to several other file formats. It would be preferable from a programming perspective if all data were stored in the same format, however. Within MATLAB alone, data can be stored in several different formats. Data may be stored in vectors/matrices; cells; cell arrays of vectors, matrices or more cells; structures; or tables, among other formats.

Even small differences in variable names such as NerveDiam versus NerveDiameter versus DiameterOfNerve are not immediately reconcilable. Without unified variable names, comparisons across datasets become very challenging.

Inconsistent variable names are not the only challenge, however. Even if variable names are identical, the values stored for that variable may be different from study to study. For example, one study might record the sex of subjects as "male" while another records it as "M" and other may code it numerically as a 1 or a 0. One study may record the locations of images from nerves as "anterior" or "posterior" while another records it as "A" or "P" and another may use a numeric distance from a landmark. Time may be stored relative to an event (e.g., time after starting an experiment) or absolute time. The latter presents several opportunities for inconsistencies as time may or may not include dates and dates may be formatted as year-month-day, day-month-year, or month-day-year formats. Months may be numeric, abbreviated strings, or full strings. Hours may be 0-24 or 0-12 with AM/PM. Within MATLAB, dates can be stored in human-readable formats such as these, or in MATLAB format that is not human readable (e.g., 24 July 2021 at 17:10:33 is represented in MATLAB as 738361.71566). Without unified data types, comparisons across datasets become very challenging.

To make the KnowMore Data Plot tool more universal we propose standardization of commonly used variable names, data formats, data types, and data units. We also recommend inclusion of key pieces of information that describe the data in the metadata. We have submitted these recommendations to SPARC and provided a copy in the supplemental section of this manuscript. This may require a significant amount of effort to convert previously uploaded datasets but should not put an exceptional burden on new studies. However, data standardization across the SPARC platform would make the data ready for much broader analysis using more sophisticated big data tools that could provide insights that are otherwise obscured or not readily accessible.

Data and Software Availability

KnowMore is fully Open Source. The latest source code is available from the KnowMore GitHub repository: xxxxx. Archieved source code as of the time of publication is available at:

License: MIT

Author Contributions

All authors contributed to preparing the manuscript. Additionally,

RQ. Develop the overall architecture of KnowMore and designed the front-end UI of KnowMore integrated into a fork of the SPARC portal.

MAS wrote the MATLAB code main.m() that tabulates data across datasets and plots these data for visual comparison. MAS wrote sections of the manuscript.

AK developed the keyword and abstract generator back-end code and contributed to the development of the Jupyter Notebook. AK wrote sections of the manuscript.

BP conceptualized the overall idea and drove the development of the software suite. BP wrote sections of the manuscript.

Competing Interests

The authors have no competing interests to disclose.

Acknowledgments

SPARC, SPARC Data Resource Center (DRC): Osparc, Pennsieve, MBF, Map-Core, DAT-CORE.

References

[i] Anita Bandrowski, Jeffrey S. Grethe, Anna Pilko, Tom Gillespie, Gabi Pine, Bhavesh Patel, Monique Surles-Zeigler, Maryann E. Martone, SPARC data structure: rationale and design of a FAIR standard for biomedical research data, bioRxiv 2021.02.10.430563; doi: https://doi.org/10.1101/2021.02.10.430563

[ii] Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

[iii] Pelot, N. A., Goldhagen, G. B., Cariello, J. E., & Grill, W. M. (2020). Quantified Morphology of the Rat Vagus Nerve (Version 4) [Data set]. Pennsieve Discover. https://doi.org/10.26275/ILB9-0E2A

[iv] Pelot, N. A., Goldhagen, G. B., Cariello, J. E., & Grill, W. M. (2020). Quantified Morphology of the Pig Vagus Nerve (Version 4) [Data set]. Pennsieve Discover. https://doi.org/10.26275/MAQ2-EII4

[v] Pelot, N. A., Ezzell, J. A., Goldhagen, G. B., Cariello, J. E., Clissold, K. A., & Grill, W. M. (2020). Quantified Morphology of the Human Vagus Nerve with Anti-Claudin-1 (Version 6) [Data set]. Pennsieve Discover. https://doi.org/10.26275/NLUU-1EWS

[vi] Pelot NA, Goldhagen GB, Cariello JE, et al. Quantified Morphology of the Cervical and Subdiaphragmatic Vagus Nerves of Human, Pig, and Rat. *Front Neurosci*. 2020;14:601479. Published 2020 Nov 4. doi:10.3389/fnins.2020.601479