Spring 5-7-2022

TROPHISH: BUILDING A GLOBAL DATABASE OF FRESHWATER TROPHIC INTERACTIONS

Jacob M. Ridgway
University of South Dakota

TROPHISH: BUILDING A GLOBAL DATABASE OF FRESHWATER TROPHIC INTERACTIONS

by

Jacob Ridgway

A Thesis Submitted in Partial Fulfillment

of the Requirements for the

University Honors Program

Department of Biology

The University of South Dakota

May 2022

The members of the Honors Thesis Committee appointed

to examine the thesis of Jacob Ridgway

find it satisfactory and recommend that it be accepted.

_____

Jeff Wesner, Ph.D.

Associate Professor of Biology

Director of the Committee

_____

Meghann Jarchow, Ph.D.

Chair & Associate Professor of Sustainability & Environment

_____

Daniel Soluk, Ph.D.

Professor of Biology

ABSTRACT

TroPhish: Building a Global Database of Freshwater Trophic Interactions

Jacob Ridgway

Director: Jeff Wesner, Ph.D.

Freshwater management and research frequently use the trophic data of freshwater fishes. Despite this fact, it is difficult to perform a simple search of dietary information for any one fish species. FishBase represents, to our knowledge, the largest compilation of freshwater dietary information to date. However, it excludes a large portion of the ecological literature due to its development taking place prior to the creation of most modern scientific search engines. Our project (TroPhish) is building upon FishBase by digitizing approximately 130 years of data from the fish predation literature. Data from the primary and grey (e.g. theses, dissertations, reports) literature were extracted, automatically scanned in through third-party software (Able2Extract), and then reorganized into a universally usable format. A total of 1123 papers were filtered, processed, and compiled to form a database with 19,893 observations of 51 variables. These observations contain data on 2808 unique dietary samples, representing 532 different species across 118 freshwater fish families from every continent fish occur. After the incorporation of FishBase's data, TroPhish will be submitted for publication in *Scientific Data*.

The TroPhish database can be accessed through its GitHub page:
https://github.com/jswesner/TroPhish

KEYWORDS: Freshwater, diet, trophic, fish, database

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACKNOWLEDGMENTS

TroPhish has taken nearly three years to develop, and its completion represents one of my greatest accomplishments to date. None of this would have been possible without a vast support network from the University of South Dakota. First and foremost, I would like to thank my Thesis Director and research mentor, Dr. Jeff Wesner. In addition to being instrumental to the development of TroPhish, his lab and care towards my success as a scientist have provided me with most of the fundamental skills and experiences that have sculpted my research trajectory. Additionally, I want to thank Melissa Beringer, Dr. Meghann Jarchow, and Dr. Daniel Soluk for giving me critical feedback on writings for both this thesis and several other scholarships/grants. I would also like to recognize Dr. Justin Pomeranz (Colorado Mesa University) for his help in developing initial code for my thesis project.

Outside of TroPhish, I would like to extend thanks to Drs. David Swanson, Mark Dixon, Ranjeet John, Will White (Oregon State University), and Jess Hopf (Oregon State University), as well as Kevin Buhl (USGS), for their mentorship roles in other research projects and academic courses. All these individuals have contributed to my development as a scientist, and I would likely not be where I am today without their support.

Background & Summary

Trophic interactions are major drivers of energy and nutrient flow within freshwater ecosystems. Alterations to these networks have historically led to trophic cascades, organism decline, or fundamental changes to population dynamics that alter freshwaters and their associated services[1,2]. Subsequently, the stability and composition of trophic interactions are powerful indicators of ecosystem health, and data surrounding them are frequently used within freshwater fisheries management. For example, population models use food web data to improve predictions of semelparous fish populations (e.g. Atlantic salmon) threatened by climate change and human exploitation[3]. Similarly, dietary data help identify the habitat and trophic relationships of vulnerable keystone species like the redspot chub to guide their management within freshwater systems[4].

The importance of dietary information in fishes has prompted hundreds of published articles reporting gut contents[5]. Nevertheless, ecological data largely consist of individual tables and graphs using variable formats[6]. As a result, it is difficult to perform a simple search of dietary information for any one fish species. The most complete compilation of fish trophic data that we are aware of is FishBase[7]. It hosts dietary information on >500 freshwater fish species, primarily based on literature searches conducted during the late 1990s[8]. These data report both the dietary items and mean trophic position of fishes globally and have been cited over a thousand times by scientists conducting trophic research.

Despite FishBase's clear value to the scientific community, its dietary data are primarily limited to data between 1980 and 2000. The reason for this is likely due to the original search for data occurring over two decades ago. Since then, technological progress and advances in the digitizing of data (e.g. meta-analyses, scientific search engines, research institutions) have surfaced unprecedented amounts of ecological information[6]. Consequently, there is still a large body of messy dietary data within scientific texts. Our project, TroPhish, builds upon the foundation of FishBase to encompass nearly 130 years of quantitative dietary data from the fish predation literature

Data from 1123 published papers, theses, dissertations, and government reports were united and filtered to form a database containing 2808 unique dietary records and 47 variables of background data for ecological context. TroPhish's dietary records represent the gut contents of 532 different fish species across 118 families and every continent except Antarctica.

Going forward, future updates are planned to reformat and incorporate FishBase's data into TroPhish. Additionally, user submission of data or errors is encouraged through our GitHub repository, https://github.com/jswesner/TroPhish, where the TroPhish database is also publicly downloadable.

Methods

Paper Collection Overview

To obtain dietary information, 1123 published papers, theses, dissertations, and government reports potentially relating to the trophic ecology of freshwater fish were gathered in three waves: 1) an initial search via Web of Science and the Minckley

Library, 2) citation chaining by searching forward and backwater citations of papers from our initial search, and 3) targeted searches on Google Scholar to expand the taxonomic scope of the initial searches (Appendix 1). Each method is described in detail below.

Initial Search

An initial literature search through Web of Science used the search terms "freshwater fish AND Diet" or "freshwater fish AND resource partitioning". In addition to Web of Science, we also searched papers from the Minckley Library (http://www.nativefishlab.net/?page_id=533), which contains >11,000 citations related to freshwater fish from both the primary and grey literature. From these searches, we excluded papers that seemed unlikely to contain dietary information by scanning the titles. This left an initial pool of 346 papers with potential dietary information, 243 from Web of Science, and 103 from the Minckley Library (Appendix 1).

Citation Chaining and Targeted Searches

Citation chaining is the process of mining references from a singular academic source[9]. This method was applied to the forward citations (i.e. papers that cited a given paper) and backward citations (i.e. papers cited by a given paper) of dozens of papers from our initial pool. Data were extracted from these papers, and geographic discrepancies were identified by plotting the locations of each study on a global map. While we had results from every continent except Antarctica, our data were heavily skewed towards North America, Europe, and South America. To expand our geographic distribution and initial pool of data, we did additional citation chaining for papers in Australia, Asia, and Africa.

In total, citation chaining yielded 475 additional papers, bringing our total to 821 (Appendix 1).

Some taxa were underrepresented at various stages of TroPhish's development. For example, despite being the 13th largest fish family[7], dietary records on Cobitidae were entirely absent from TroPhish. Others, like Mormyridae and Rivulidae, were represented but comparatively infrequent to their actual diversity. To address this, we sourced the remaining 302 papers from 28 targeted literature searches through Google Scholar (Appendix 1).

Data Digitization

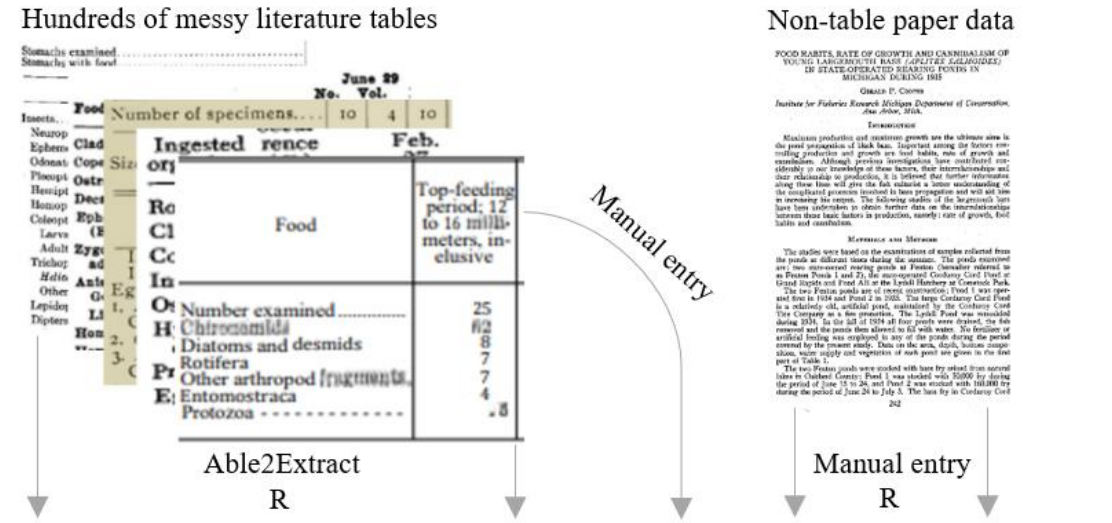
All 1123 papers were read manually and filtered to capture those containing data on the weight, abundance, or volume of individual prey taxa in fish diets (Appendix 1). These data had to be reported either as a proportion of the total diet or as a raw value in the form of a table. Tables from papers meeting the above criteria were converted to .csv files using Able2Extract (Version 6.0, Investintech), a program specialized in converting tables in PDFs to Microsoft Excel files. Data able to be digitalized through this software were not retained.

Database formation

Digitized tables were then individually rearranged into a common format in Excel where tables largely followed tidy principles[10], with each column representing a single variable, each observation representing a row, and each observational unit representing a table (Fig. 1).

During this process, 24 variables worth of additional data were manually extracted to provide ecological contexts for six general areas (Fig. 1):

1. Predator-prey taxonomy and organism ecology
2. Temporal information
3. Geographic information
4. Habitat information
5. Sampling methodology
6. Data source



Prey item	Measurement	Measurement Type	Min length	Max length	Fish	
Chironomidae	62	relative count	12	16	Catostomus commersonii	…
Diatoms and desmids	8	relative count	12	16	Catostomus commersonii	…
Rotifera	7	relative count	12	16	Catostomus commersonii	…
Arthropod fragments	7	relative count	12	16	Catostomus commersonii	…
Entomostraca	4	relative count	12	16	Catostomus commersonii	…
Protozoa	0.5	relative count	12	16	Catostomus commersonii	…

47 other variables

Figure 1. A visualization of TroPhish's data tidying process. The literature table lacks variables that would be needed to compare data (e.g. Max length) and instead treats values (e.g. 12-16 mm) like variables. Our tidying process adds the missing variables and reorganizes the data into universal format that can be stacked into a single database (TroPhish) and analyzed.

Most literature tables contained multiple dietary records. Instead of creating separate files or repeating prey items in Excel for each unique sample, prey taxa and their quantitative measurements were transposed and treated as variables and values, respectively. Tables, now all commonly formatted, were then compiled into a single database using R[11], with the *mutate()* and *gather()* functions from the *tidyverse*[12] package (Fig. 1).

Post-Hoc Processing & Verification

Naming conventions for fish and prey taxa varied widely among papers (e.g., *Chironomidae* was written in at least 15 different ways: *chiro, chiros, chiron, chironomid,* ect.; *Lepomis macrochirus* was written variously as *L. macrochirus, Bluegill, L.m.,* etc). Therefore, we used a mix of manual and automated methods to assign formal taxonomic names to fish and prey.

For prey items, we assigned taxonomic information by matching names in TroPhish with taxonomic information from the National Center for Biological Infrastructure using the R package *taxize*[13]. Names without a match, due to the variations shown above, were then split and taxonomic information was added manually. The manual and automated information was then merged and checked manually for misspellings or misidentified taxa, and then appended to the original dataset. This preserved the variations in spellings from the original papers while also adding the formal taxonomic information.

For fish, we did a similar procedure as above but used the R package *rfishbase*[14] as an initial source of taxonomic information. In most cases, assigning taxonomic information was straightforward. In cases where colloquialisms or odd abbreviations were used, we consulted the text of the original paper to see if the full names were given. If they were

not, then we searched Google Scholar for other papers that used the same spellings. If that failed, then we entered the taxon as “unknown”. Upstream taxonomic information (e.g. order, subclass, family, etc.) was obtained using the R packages *taxize* and *rfishbase* (e.g. *Lepomis macrochirus* can be converted to species: *Lepomis macrochirus,* Genus: *Lepomis,* Family: *Centrarchidae*, etc.).

Some papers also included life-stage information, such as larvae, pupae, and adults. These also had various spellings (e.g., *adults* was written as *adult, a., ad., imago,* etc.). To consolidate this information, we used the *separate()* and *distinct()* functions in *tidyverse* to obtain a list of all possible spellings of life-stage information. We then formalized all names manually (e.g., *adults* and *imago* would all be *adults*) and added them to the original dataset.

Geographic coordinates (latitude and longitude) were geocoded from all unique sampling sites obtained during our initial data extraction process. This was done with the function *mutate_geocode*() from the R package *ggmap*[15]. As with taxonomic names, not all locations were discoverable. For those, we did manual searches on Google Maps or used other information in the paper to estimate latitude and longitude. Coordinates were then plotted on a map for both visualization purposes (Fig. 2) and to find obvious geocoding errors. All data are accurate to the smallest scale we could efficiently extract given what was stated in each paper. Typically, accuracy fell to the level of the water body where sampling occurred. However, some sites reduced our precision to large swaths of territory (e.g. state, country, subregion).

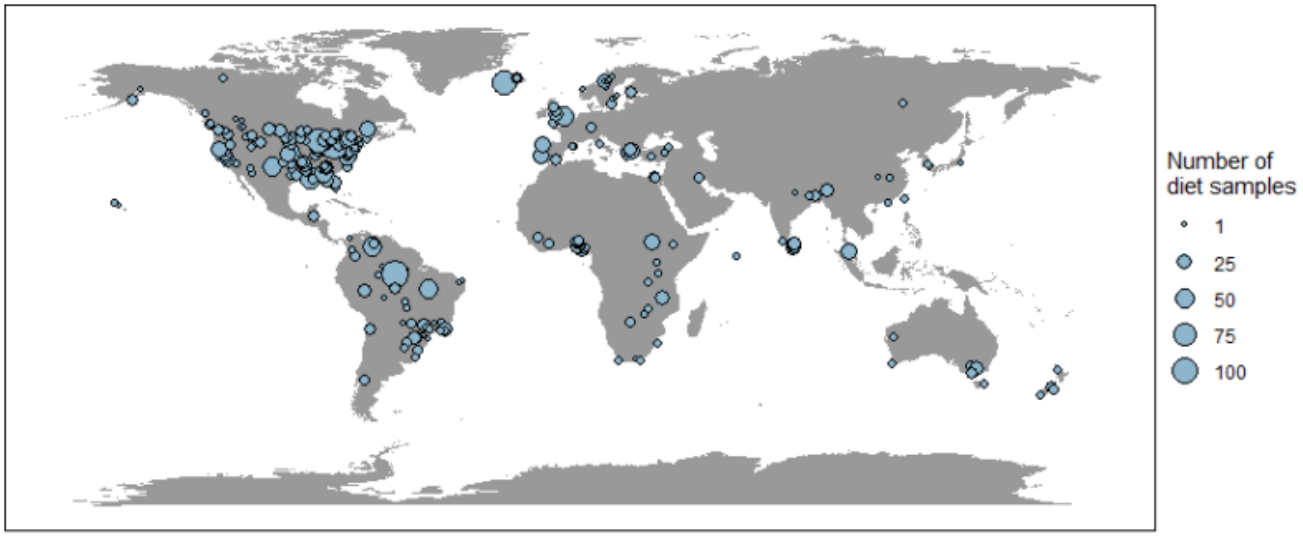
Data Records

Table 1. Literature parameters of the TroPhish database.

Tables	*Papers*	*Authors*	*Journals*	*Years*
518	267	>250	164	132

The TroPhish database contains 19,893 observations of 51 variables from hundreds of tables, papers, authors, and journals (Table 1). These observations represent 2808 unique dietary samples from 532 different species and 118 freshwater fish families from every freshwater fish-habitable continent (Fig. 2).

Figure 2. A geographic visualization of the 2808 unique freshwater fish dietary samples in our database, based on data taken from 267 papers.



Against the actual diversity of 10,723 species across 256 families[7], our project currently represents ~5% of freshwater fish species and 46% of all freshwater fish families. These are parameters comparable to international efforts like FishBase, which represents the trophic information of 567 species across 111 families[7].

Most diet studies are targeted at socially or economically important and evolutionarily unique species[16]. Thus, as expected, fish families in those categories (e.g. gamefish like

perches, trout, and salmon) were overrepresented in TroPhish, relative to their taxonomic diversity (Fig. 3). To give an instance, while Salmonidae represent 1.3% of all freshwater fish species[7], they encompass 2.3% of TroPhish's fish species (Fig. 3). Regardless, the literature still maintained adequate representation of major fish families, such as Cichlids, Cyprinids, and Characids (Fig. 3).

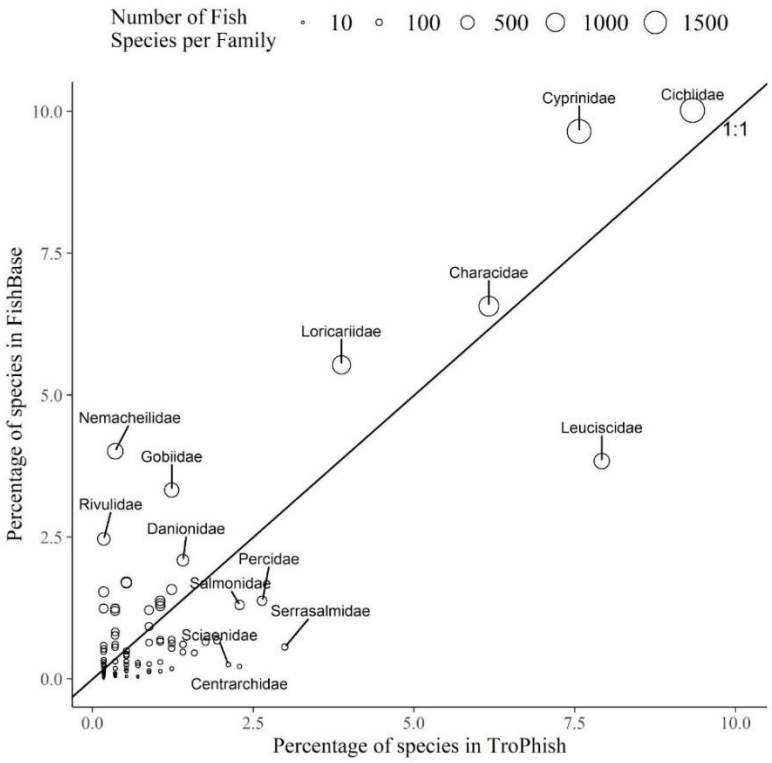


Figure 3. Comparison of the species coverage in TroPhish versus global species in FishBase. Dot size represents the total number of species in a given family globally via FishBase. The x-axis represents the proportion of species in TroPhish that come from a given family (e.g., most species come from Cichlidae and Cyprinidae and fewer from Centrarchidae). The y-axis represents the same proportion globally. Values below the 1:1 line indicate proportionally more species in TroPhish. Values above it indicate proportionally more species in FishBase. For visual clarity, labels are limited to families with >2% of species in at least one database.

Technical Validation

Literature validation

Comparing the forward and backward citations of heavily cited papers during citation chaining to our database provided a crude method of approximating TroPhish's representation of the literature. Before the final addition of 321 papers, our database averaged 53% of commonly cited literature. This average varied over time as papers were

added to TroPhish (Fig. 4). Nevertheless, our dataset considered multiple citations from nearly all papers throughout our checks (Fig. 4), suggesting broad coverage.

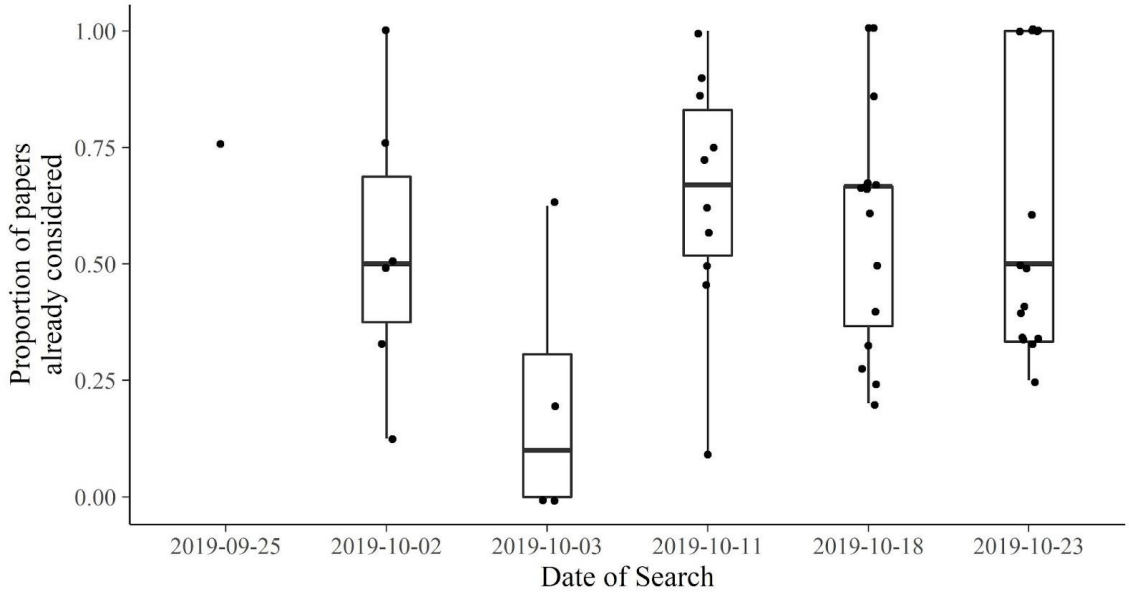


Figure 4. The proportion of papers in backward or forward citations that had already been considered by the date of the search. Each dot is an individual diet paper. The y-axis is the proportion of papers cited in that paper that had already been considered for inclusion in our database. Papers not considered were added to the pool for later consideration. On average, more than 53% of papers found had already been considered, indicating good coverage of diet papers in TroPhish.

Scanning and manual entry validation

Data on relative dietary measurements (e.g. percent measurements of volume, mass, abundance) were individually totaled to find scanning errors that would deviate the data from ~100%. Non-relative dietary data (e.g. non-percent measurements of volume, mass, abundance, etc.) were manually compared to their values within the literature to detect additional PDF to Excel conversion errors. Data across all remaining variables were scanned for entry errors through a mixture of automatic code in R and proofreading.

Additionally, all quantitative data were plotted to find extreme deviations likely to be representative of an error. Tables containing errors were corrected within R or Excel to match what was reported by their associated papers. Though only a few papers, data still containing errors past this point were not retained.

Future updates

At least one large future update is planned to incorporate the dietary data of FishBase. Additionally, users can make suggestions on GitHub to send in data and report errors through push requests (see *Usage Notes* for more details), which will then be incorporated into TroPhish through periodic updates.

Usage Notes

Obtaining the dataset and user submission

The TroPhish database is available to download in multiple formats on our GitHub page, https://github.com/jswesner/TroPhish. Users are encouraged to submit errors or additional data through GitHub's fork and pull model (https://docs.github.com/en/pull-requests/collaborating-with-pull-requests/getting-started/about-collaborative-development-models). Submitted data must be aligned with the format of TroPhish. Excel files of the literature tables comprising TroPhish are provided for easy error submission and formatting reference.

Use case

To demonstrate how TroPhish can be used to summarize ecologically important data, we plotted the proportion of diets that consisted of terrestrial prey for two common fish families (Leuciscidae/Salmonidae). This question is routinely assessed in individual

dietary studies and meta-analyses[17] due to the importance of terrestrial subsidies to the function of stream ecosystems[18]. As shown in Figure 5, the data in TroPhish suggest that terrestrial prey make up less than 25% of fish diets in most genera, but with large variation among genera. Additionally, the plot suggests areas that warrant future studies. For example, the genera Notropis, Macrhybopsis, Onchorhynchus, and Salmo are relatively well-studied, while Squalius, Gila, Prosopium, and Thymallus are not.

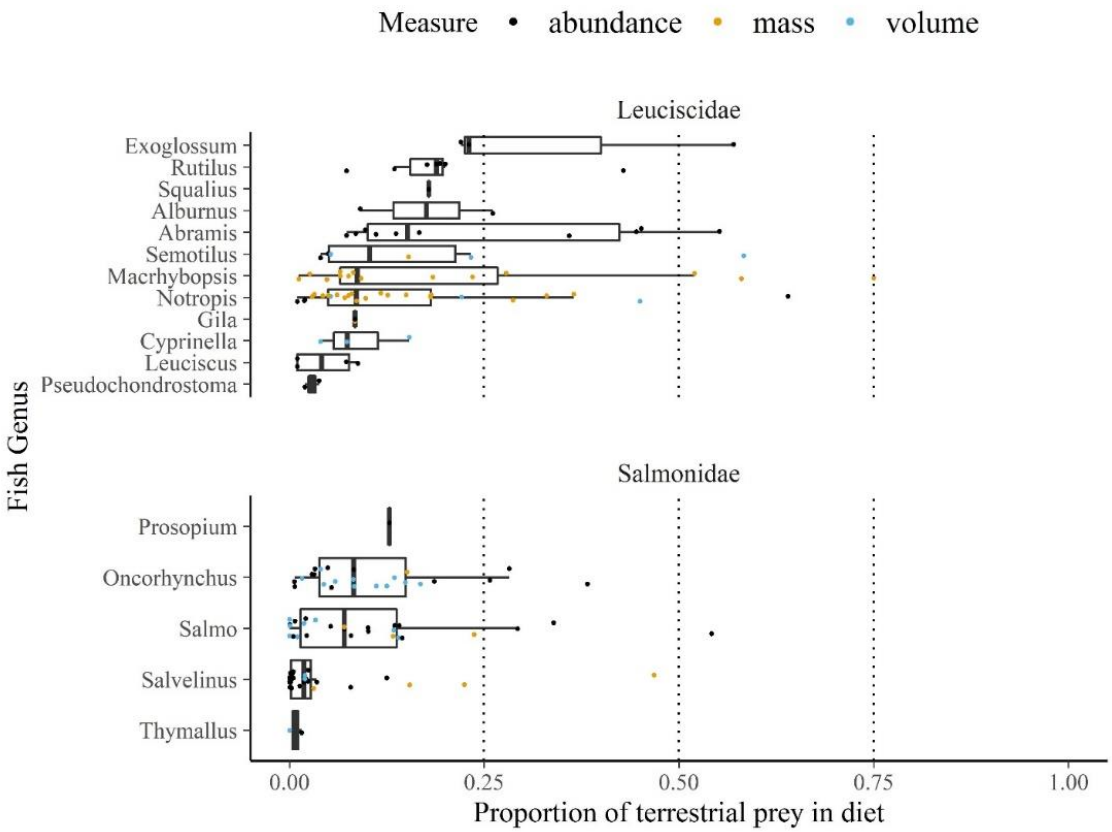


Figure 5. The proportion of terrestrial prey consumed by the genera of Salmonidae and Leuciscidae represented by TroPhish. Each dot represents a unique dietary sample. The y-axis represents the relative abundance (black), relative mass (yellow), or relative volume (blue) of terrestrial prey contained within the diets of each genus (x-axis).

References

1. Helfield, J. M. & Naiman, R. J. Keystone Interactions: Salmon and Bear in Riparian Forests of Alaska. *Ecosystems* **9**, 167–180 (2006).

2. Power, M. Effects of fish in river foodwebs. *Science* **250**, 811–814 (1990).

3. Woodward, G. *et al.* Using Food Webs and Metabolic Theory to Monitor, Model, and Manage Atlantic Salmon—A Keystone Species Under Threat. *Front. Ecol. Evol.* **9**, 675261 (2021).

4. Rodger, A. W. & Starks, T. A. Ontogenetic Diet Shift, Feeding Ecology, and Trophic Niches of the Redspot Chub (Cypriniformes: Cyprinidae: Nocomis asper). *Ichthyology & Herpetology* **109**, (2021).

5. Hyslop, E. J. Stomach contents analysis-a review of methods and their application. *J Fish Biology* **17**, 411–429 (1980).

6. Hampton, S. E. *et al.* Big data and the future of ecology. *Frontiers in Ecology and the Environment* **11**, 156–162 (2013).

7. Froese, R. & Pauly, D. FishBase. (2010).

8. Palomares, M. L. & Sa-a, P. The DIET Table. (2000).

9. Cribbin, T. F. Citation chain aggregation: an interaction model to support citation cycling. in *Proceedings of the 20th ACM international conference on Information and knowledge management - CIKM '11* 2149 (ACM Press, 2011). doi:10.1145/2063576.2063913.

10. Wickham, H. Tidy Data. *J. Stat. Softw.* **59**, 1–23 (2014).

11. Ihaka, R. & Gentleman, R. R: a language for data analysis and graphics. *J Comput Graph Stat* **5**, 299–314 (1996).

12. Wickham, H. *et al.* Welcome to the Tidyverse. *JOSS* **4**, 1686 (2019).

13. Scott A. Chamberlain & Eduard Szöcs. Taxize Source Code And Program Files. (2013) doi:10.5281/ZENODO.7097.

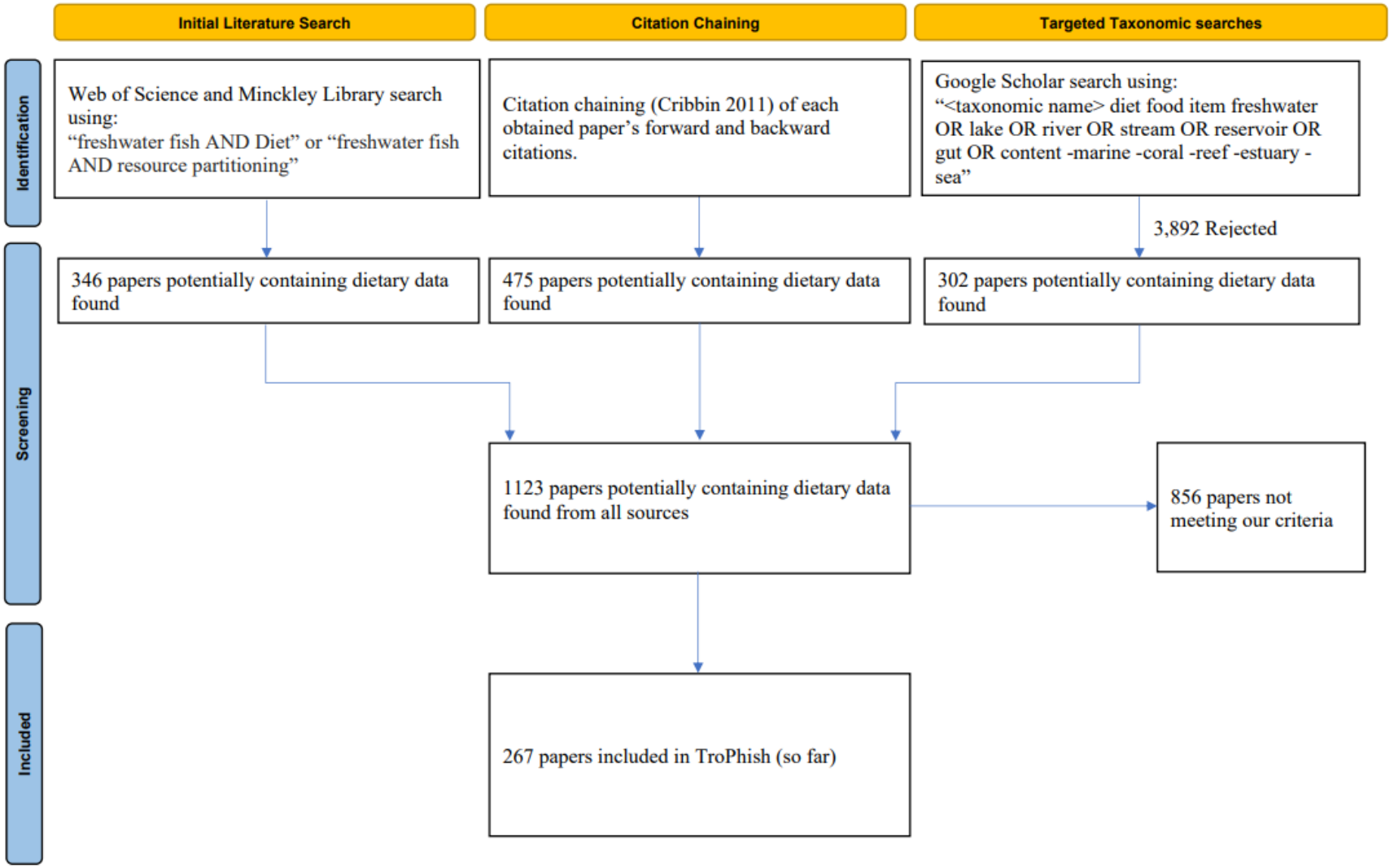
14. Boettiger, C., Lang, D. T. & Wainwright, P. C. rfishbase: exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology* **81**, 2030–2039 (2012).

15. Kahle, D. & Wickham, H. ggmap: Spatial Visualization with ggplot2. *The R Journal* **5**, 144 (2013).

16. Troudet, J., Grandcolas, P., Blin, A., Vignes-Lebbe, R. & Legendre, F. Taxonomic bias in biodiversity data and societal preferences. *Sci Rep* **7**, 9132 (2017).

17. Sullivan, M. L., Zhang, Y. & Bonner, T. H. Terrestrial subsidies in the diets of stream fishes of the USA: comparisons among taxa and morphology. *Mar. Freshwater Res.* **63**, 409 (2012).

18. Nakano, S. & Murakami, M. Reciprocal subsidies: Dynamic interdependence between terrestrial and aquatic food webs. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 166–170 (2001).



Appendix 1. PRISMA flowchart of TroPhish's data collection. A total of 1123 papers potential containing freshwater dietary data were extracted from Web of Science (243), the Minckley Library (103), Google Scholar (302), and citation chaining (475). Of this initial pool, only 267 were included in TroPhish.