SQL databases

An introduction

			3	11798	Carnivora	Felidae	Felis	sp.	S. Kuemmell	C. Linnaeus	1758	10883	Lufeng	China	Yunnan
			4	11798	Carnivora	Felidae	Metailurus	sp.	H. O'Regan	Zdansky	1924	10883	Lufeng	China	Yunnan
			5	11803	Carnivora	Felidae	Panthera	pardus	W. Clyde	C. Linnaeus	1758	10604	Tabun Cave Level C & D	Israel	Mount Carmel
			6	13066	Carnivora	Felidae	Panthera	onca	M. Uhen	C. Linnaeus	1758	10272	Ladds Quarry	United States	Georgia
			7	13066	Carnivora	Felidae	Lynx	rufus	M. Uhen	Schreber	1777	10272	Ladds Quarry	United States	Georgia
			8	13066	Carnivora	Felidae	Miracinonyx	inexpectatus	M. Uhen	E. D. Cope	1895	10272	Ladds Quarry	United States	Georgia
			9	13066	Carnivora	Felidae	Puma	concolor	J. Alroy	Linnaeus	1771	10272	Ladds Quarry	United States	Georgia
			10	13066	Carnivora	Felidae	Felis	sp.	M. Uhen	C. Linnaeus	1758	10272	Ladds Quarry	United States	Georgia

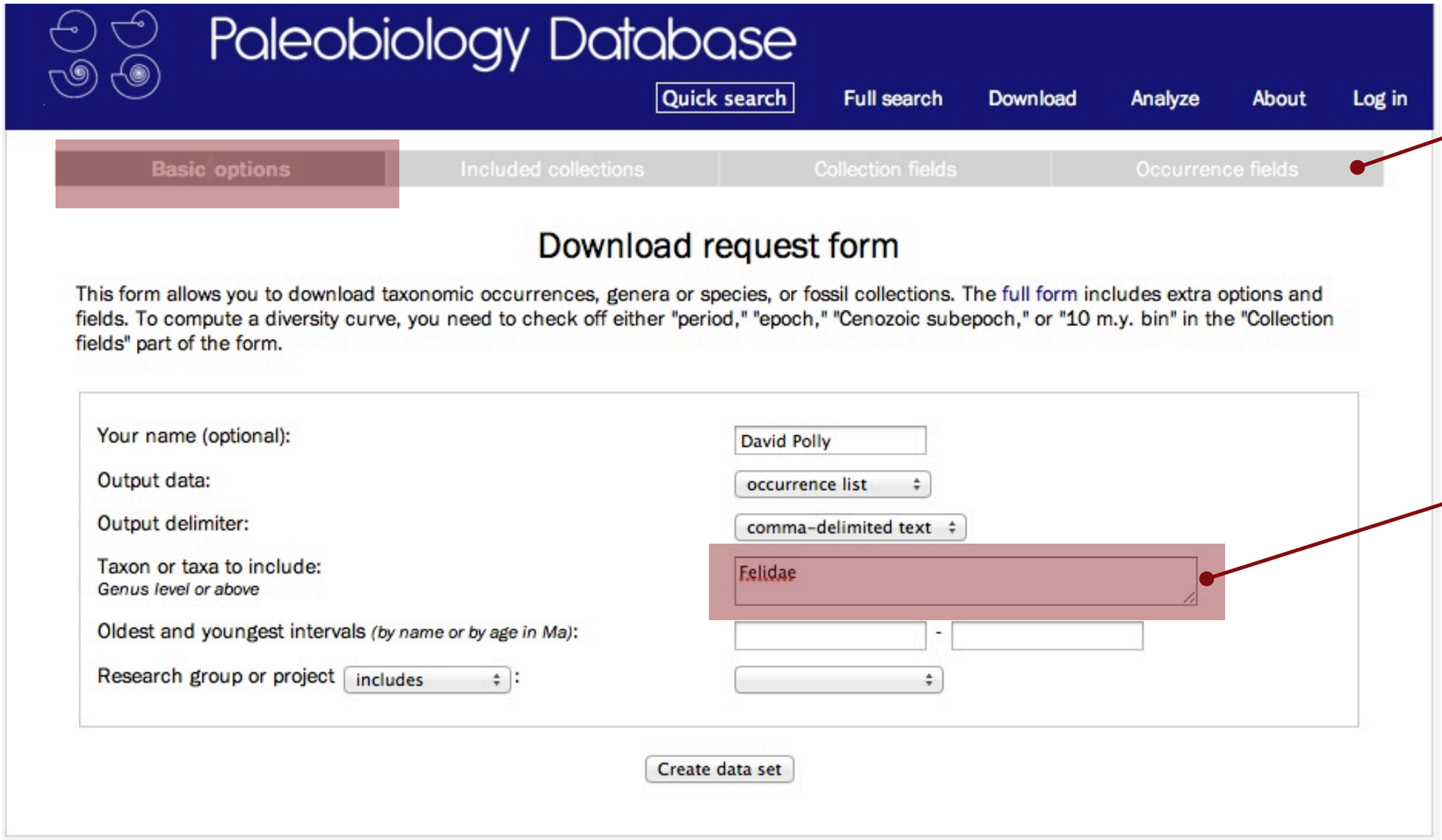
Today's Session

1. Selecting data from Paleobiology Database
2. Downloading and reformatting PaleoDB data
3. Setting up AMP servers and understanding them
4. Understanding a database server
5. Using phpMyAdmin client to operate mySQL
6. Importing data into mySQL
7. Break
8. Querying data using the SQL language
9. Assignment

The Paleobiology Database

- Community effort to record fossil occurrence data from published literature for use by the community
- Started in 1998 by project funded through the National Center for Ecological Analysis and Synthesis
- John Alroy of Macquarie University coordinates the PaleoDB
- Data are peer-reviewed and entered by scientists working on volunteer basis
- Data are extensive
- Data are incomplete: not all time, localities, or taxa are represented and there may be many omissions and errors
- Other, more specialized occurrence databases include:
  - NOW (Neogene Old World Database, coordinated by Mikael Fortelius)
  - FaunaMap (Database of mammal occurrences in the Pleistocene of North America, coordinated by the Illinois State Museum)
  - MioMap (Database of mammal occurrences in the Miocene of North America, coordinated by Tony Barnosky)
- The above are similar kinds of community efforts that focus on particular times and taxa, with similar strengths and weaknesses
- Always acknowledge or cite the database, the data enterers, and the authors of the original literature from which the data was derived

PaleoDB Download: 1. Basic Options Page

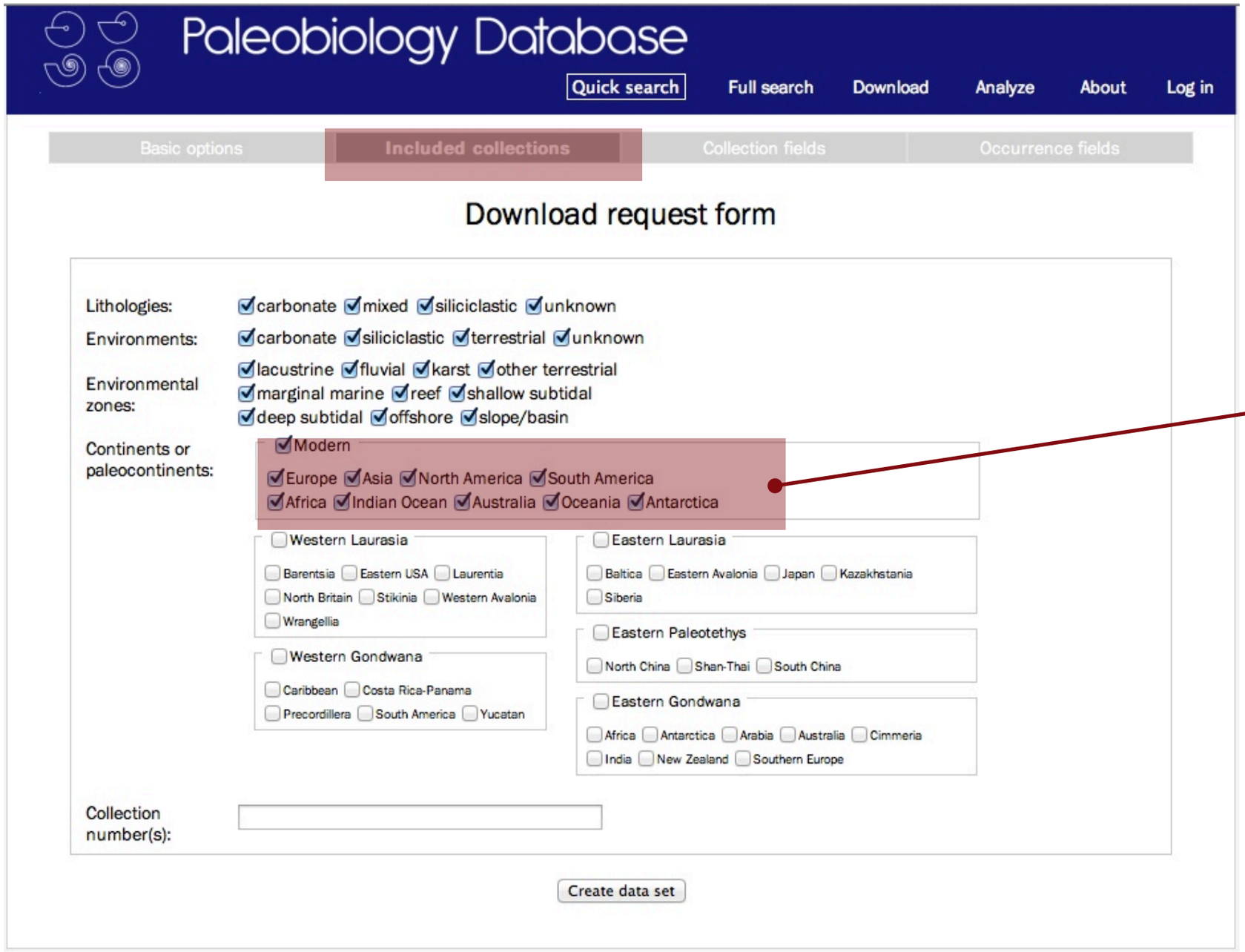


Move between tabs here

Taxonomic name here

Screen capture from Paleobiology Database (http://pbdb.org/)

PaleoDB Download: 2. Included Collections Page

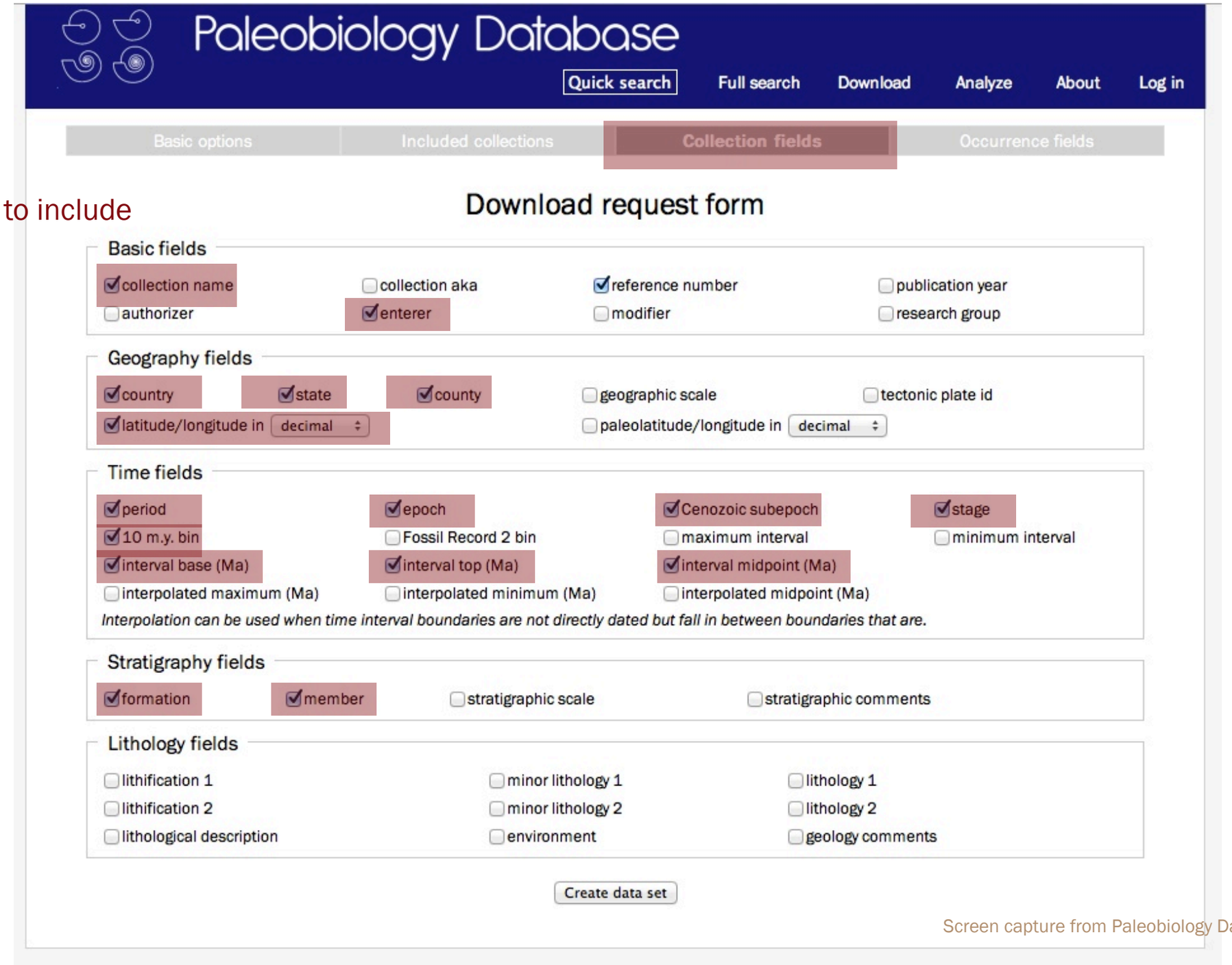


Limit search by continent here if you want

Screen capture from Paleobiology Database (http://pbdb.org/)

PaleoDB Download: 3. Collection Fields Page

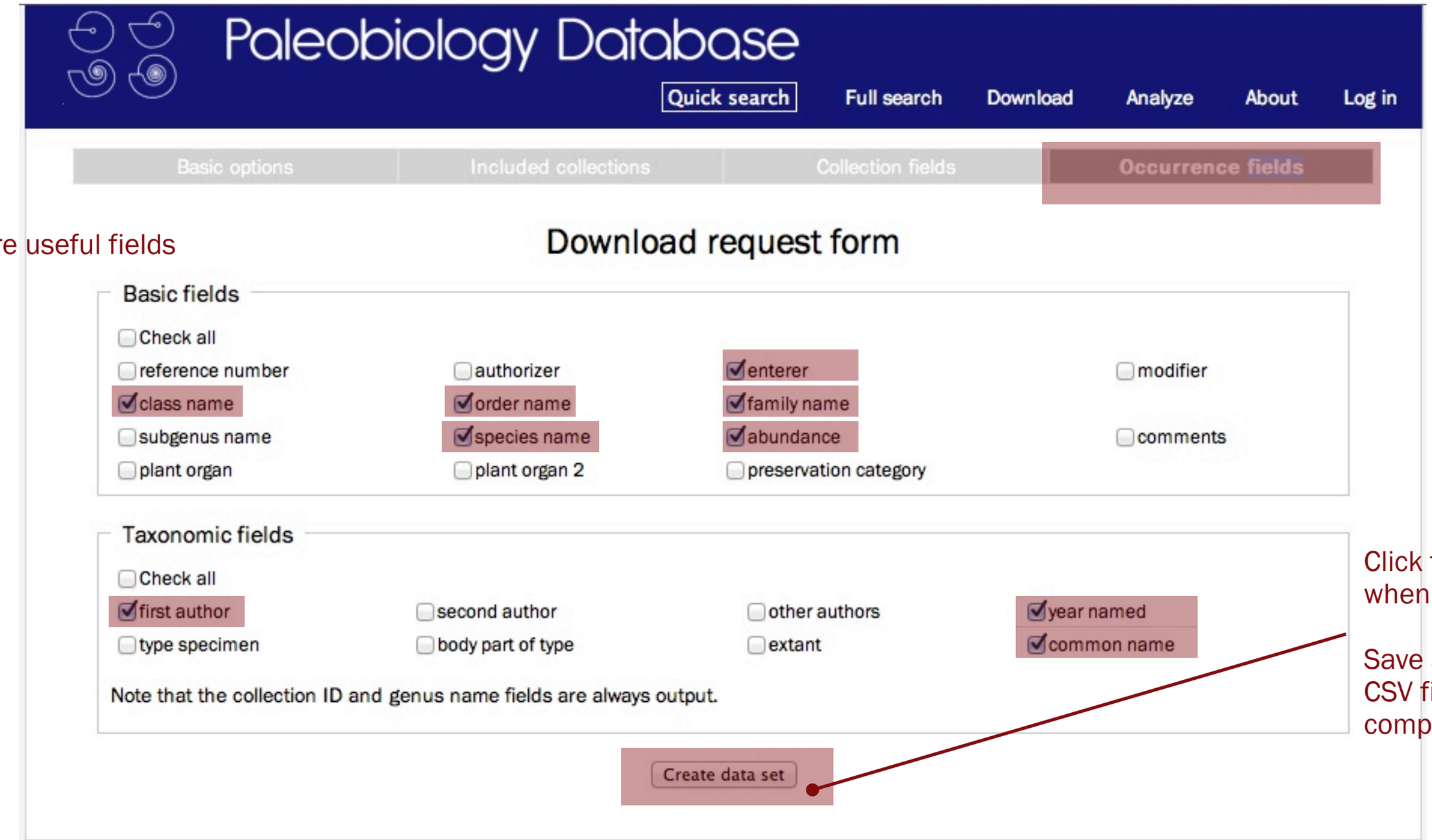
Useful fields to include



Screen capture from Paleobiology Database (http://pbdb.org/)

PaleoDB Download: 4. Occurrence Fields Page

More useful fields



Click to search when ready

Save all three CSV files to your computer

Screen capture from Paleobiology Database (http://pbdb.org/)

Remember to give credit...

- ✳ to the Paleobiology Database
- ✳ to the contributors of the data
- ✳ to the original authors of the data

Your download is large. If you intend to use the data in a publication, we ask you to:

- Notify some of the major contributors about your research.
- Acknowledge the major contributors by name in your paper.
- Cite the relevant Online Systematics Archive.
- If none are relevant, give a generic citation to the Paleobiology Database including the date of access.
- Join the PaleoDB and contribute more data.
- At a bare minimum, request a publication number once your paper is accepted.

The major contributors to this data set were:

- John Alroy (44.7%)
- Kay Behrensmeyer (23.0%)
- Alan Turner (19.2%)

Each percentage is based on the estimated amount of time spent entering data by the authorizer and associated data enterers. The grand total is 748 hours (equivalent to 4.5 work months). The estimate is based on (1) date stamps for keystroked collections and (2) counts of published references used to document uploaded collections.

Additional contributors include Chris Bell, Matt Carrano, Will Clyde, Darin Croft, Emmanuel Fara, Jason Head, Denne Reed, Mark Uhen, Xiaoming Wang, and Lars van den Hoek Ostende.

If you agree to these terms, click here to access the data.

Here are the publications that yielded the most records in this data set, in order of importance. Please cite some or all of them in any publication based on it. If possible, please also include the DavidPolly-refs.csv file as online supplementary information.

M. D. Leakey and J. M. Harris. 1987. Laetoli: a Pliocene Site in Northern Tanzania. *Clarendon Press, Oxford, Great Britain* [17 collections, 45 occurrences]

M. Bon, G. Piccoli, and B. Sala. 1991. I giacimenti Quaternari di vertebrati fossili nell'Italia Nord-Orientale. *Memorie degli Instituti di Geologia e Mineralogia dell'Universita di Padova* **47**:185-231 [11 collections, 27 occurrences]

E. I. Belyaeva. 1948. *Catalogue of Tertiary Fossil Sites of the Land Mammals in the U.S.S.R.* [10 collections, 26 occurrences]

W. von Koenigswald and W.-D. Heinrich. 1999. Mittelpleistozane Saugetierfaunen aus Mitteleuropa - der Versuch einer biostratigraphischen Zuordnung. *Kaupia* **9**:53-112 [7 collections, 29 occurrences]

E. Heintz, C. Guerin, R. Martin and F. Prat. 1974. Principaux gisements villafranchiens de France: Listes fauniques et biostratigraphie. *Memoires du Bureau de Recherches Geologiques et Minieres* **78(1)**:169-182 [7 collections, 25 occurrences]

G. D. Koufos. 2001. The Villafranchian mammalian faunas and biochronology of Greece. *Bollettino della Societa Paleontologica Italiana* **40(2)**:217-223 [10 collections, 15 occurrences]

D. Janossy. 1986. Pleistocene vertebrate faunas of Hungary. *Developments in Palaeontology and Stratigraphy, 8. Elsevier, Amsterdam* 1-208 [9 collections, 16 occurrences]

M. L. Cassiliano. 1999. Biostratigraphy of Blancan and Irvingtonian mammals in the Fish Creek-Vallecito Creek section, southern California, and a review of the Blancan-Irvingtonian boundary. *Journal of Vertebrate Paleontology* **19(1)**:169-186 [12 collections, 10 occurrences]

AMP: Apache, mySQL, PHP

MAMP: One-click-solution for setting up your personal webserver

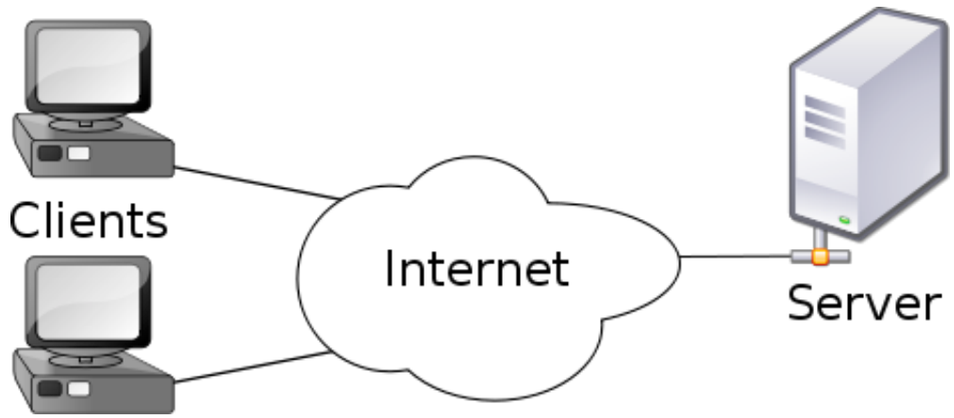
This installations installs the Apache webserver, the PHP scripting language, and the mySQL database on your computer:

- **Apache:** runs in the background on your computer and waits for requests from web browsers, sending back the requested files (http://localhost/ to see the site now running on your laptop)

- **mySQL:** a relational database server that runs in the background of your computer and waits for requests from SQL clients, returning to them the requested data

- **PHP:** a computer language that was specially designed for use on internet servers, making it easy to extract customized data from databases and to reform it into web pages

Client-server model



A method of distributing access to a program

Server is the main program that handles information, calculations, data analysis, etc.

Client is the computer or program that asks for and receives information from the server

Web: client is your web browser, server is a web server application like Apache.

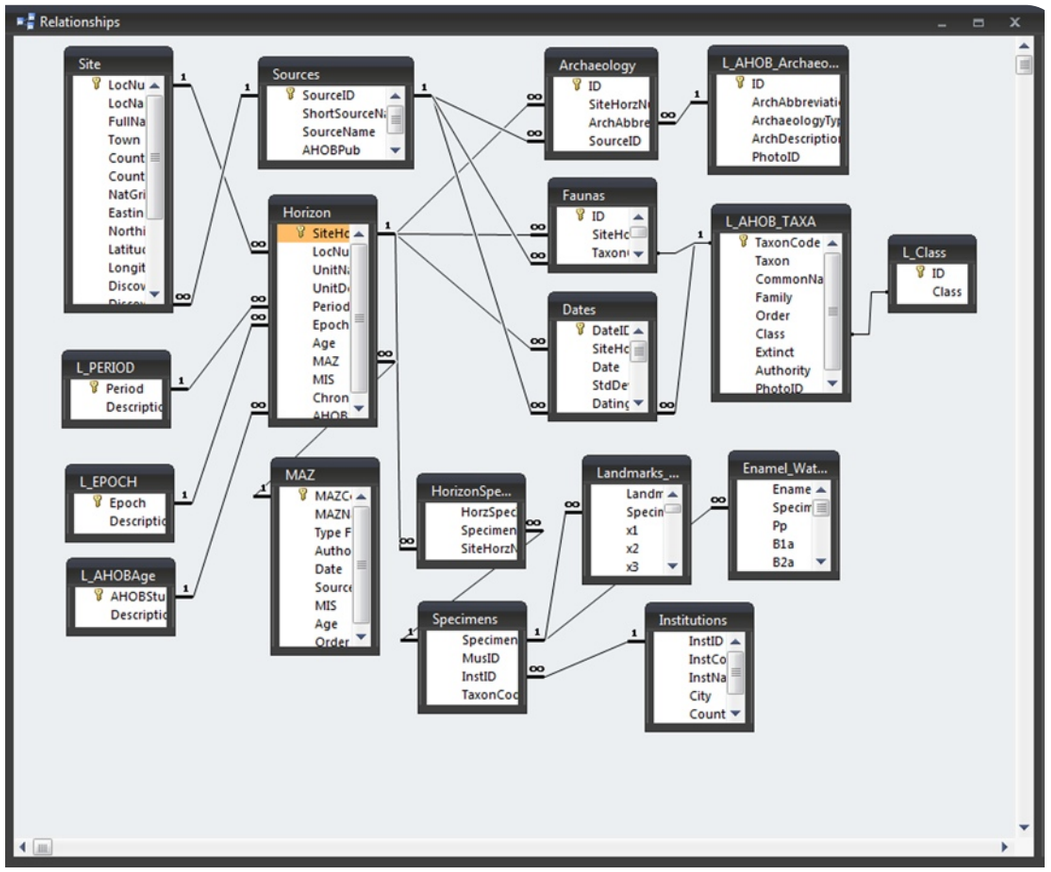
Databases: client is the program you use to query database, server is mySQL installed on a machine somewhere

iPhone Apps: client is the app installed on iPhone, servers is (usually) a web server like Apache

Advantage of client server model for databases: you can access data from MANY different programs without reformatting your data.

What is a relational SQL database?

- Software that stores and processes data
- Relational aspect refers to the ability to keep data in many tables that are linked by particular variables (e.g., "specimen" table may list of data about specimens in a museum, "taxon" table may contain information about species, the link allows users of the specimen table to find out about the species).
- SQL (Structured Query Language) is a simple language or syntax for issuing instructions to the database software
- Databases are useful for managing large amounts of data and for reformatting or collating them



from Polly & Stringer, 2011

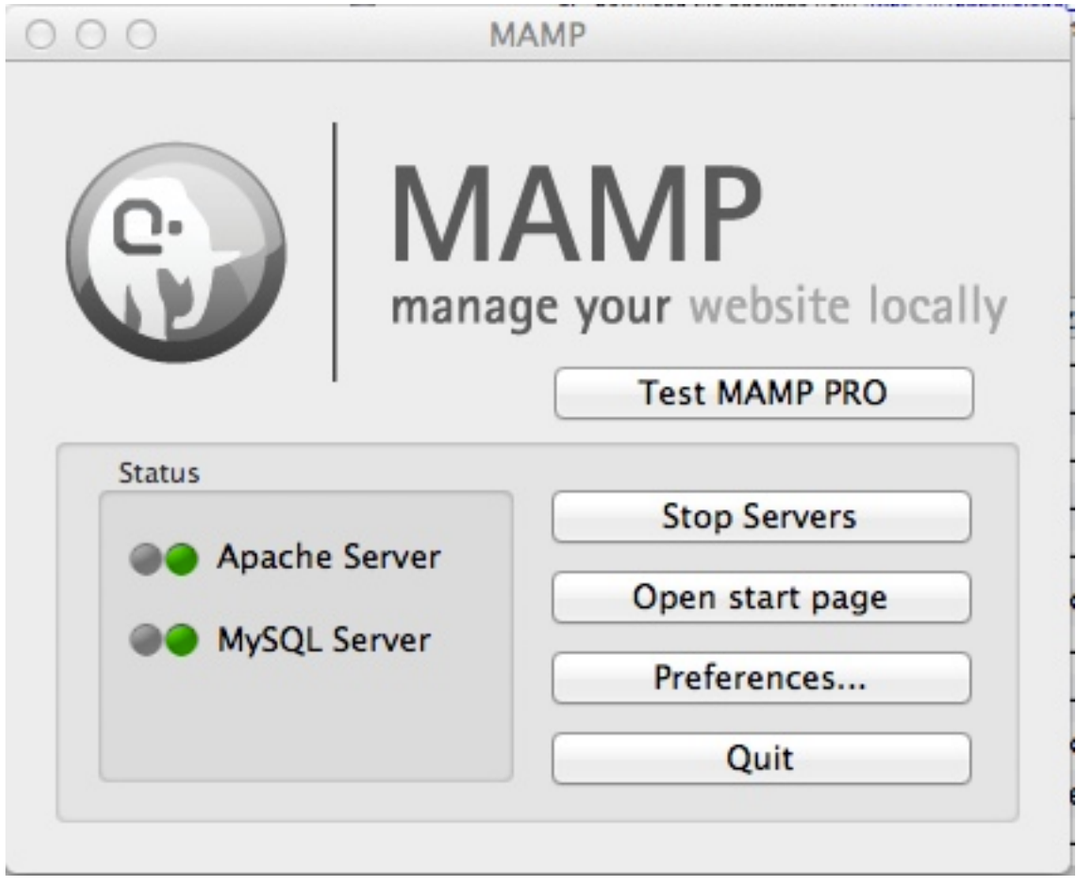
Graphic summary of relationships in a complex database with eighteen tables

mySQL database

- A free, open-source relational database (now owned by Oracle)
- SQL compliant
- Built on a client-server model
- Fast, can handle huge data sets
- Cross-platform
- Specially designed to interface with web applications using scripts (PHP, PERL, Python, Java, etc.)

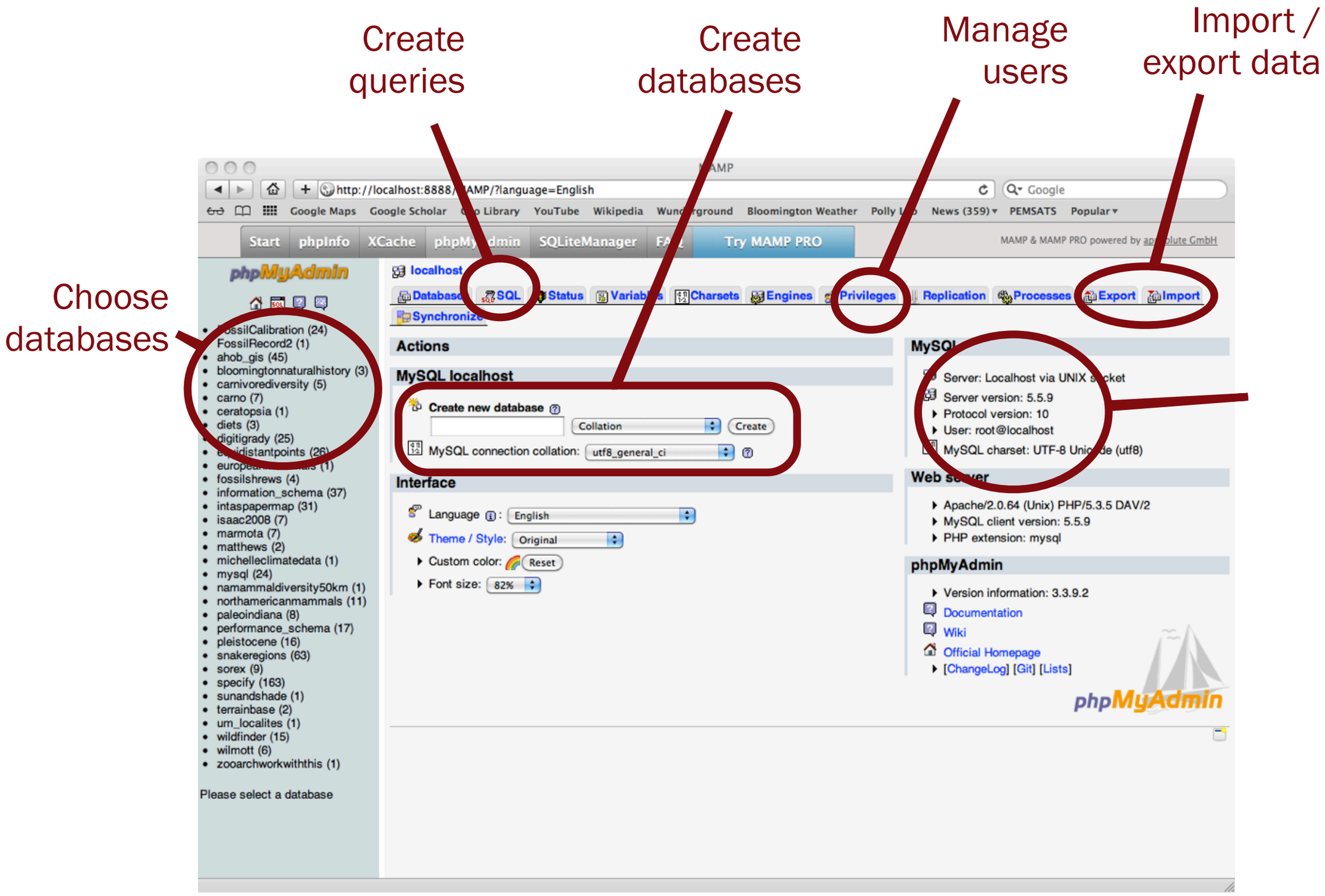
Start the mySQL Database Server and phpMyAdmin

- “Start Servers” starts both mySQL database and Apache web server
- Always stop servers when you are not using them
- Start page opens the PHPMyAdmin control interface

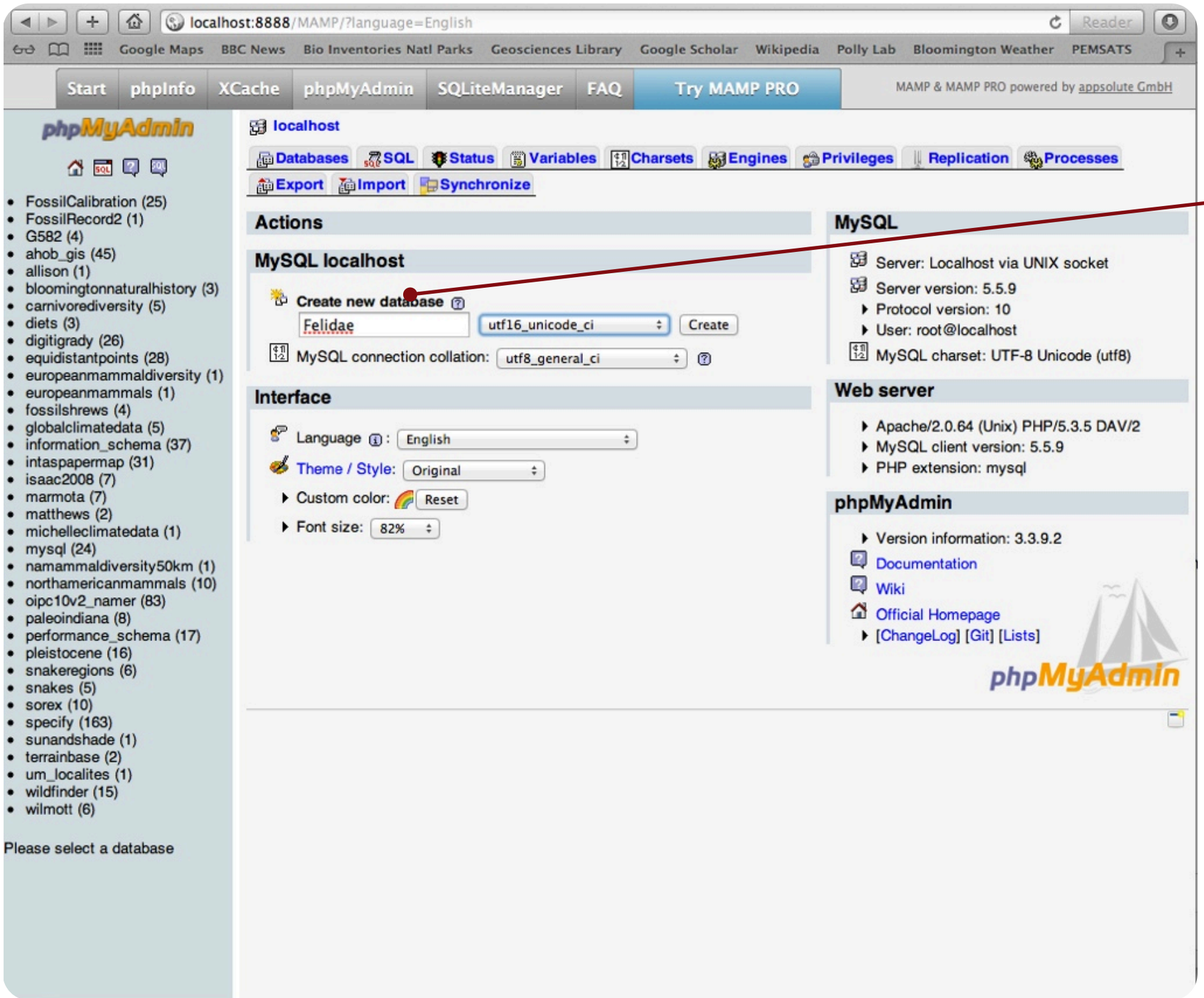


Mac OS X version. WAMP control looks different.

phpMyAdmin layout



Creating a database in phpMyAdmin



Create new database form (on home page)

Name your database and choose utf16_unicode_ci from the Collation dropdown menu

Press Create when ready

Reformat PaleoDB data for import into mySQL

Using Excel (or other spreadsheet):

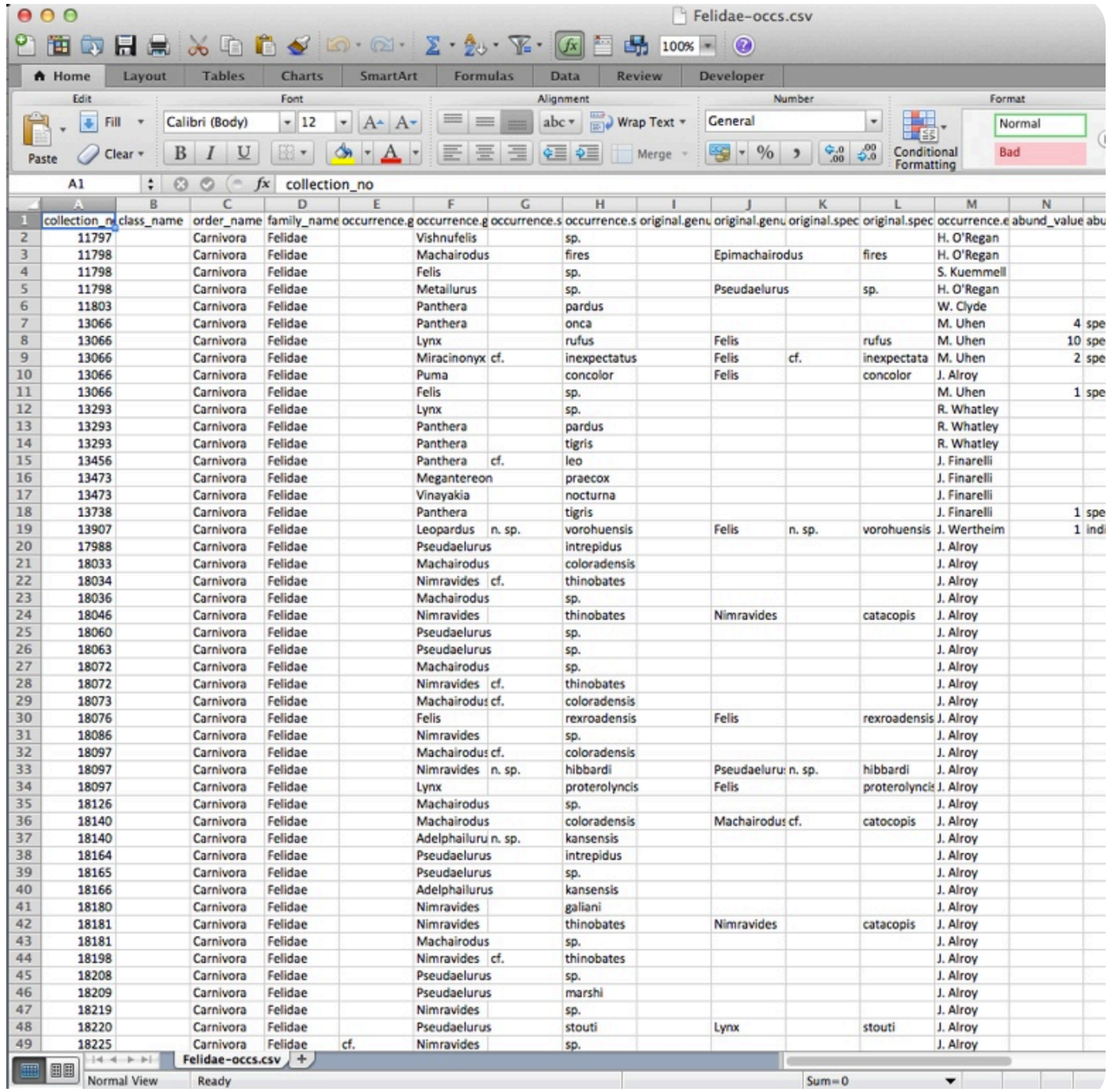
Delete columns that do not seem well-populated or useful.

Rename columns with a short and intelligible label. For example:

"order_name"-> "Order"
"occurrence.genus_name" -> "Genus"
"latdec" -> "Latitude"
"longdec" -> "Longitude"
"interval_base" -> "MaxAge"
"interval_base" -> "MinAge"
"interval_midpoint" -> "MedAge"

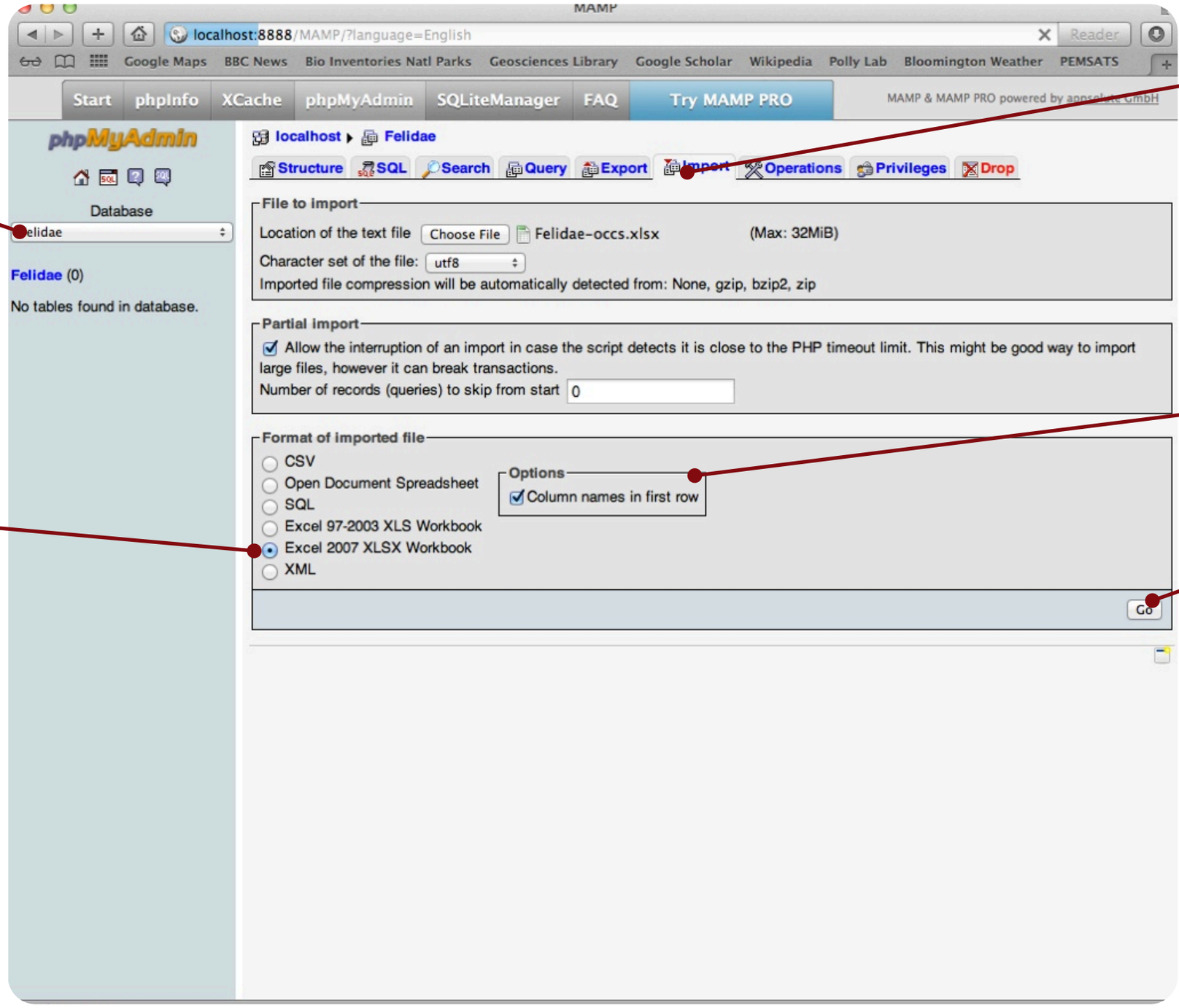
Do not use the same column name twice.

Save as an Excel workbook instead of comma-delimited file.



collection_no	class_name	order_name	family_name	occurrence.g	occurrence.g	occurrence.s	occurrence.s	original.genu	original.genu	original.spec	original.spec	occurrence.e	abund_value	abu
11797		Carnivora	Felidae		Vishnufelis		sp.					H. O'Regan		
11798		Carnivora	Felidae		Machairodus		fires		Epimachairodus		fires	H. O'Regan		
11798		Carnivora	Felidae		Felis		sp.					S. Kuemmell		
11798		Carnivora	Felidae		Metailurus		sp.		Pseudaelurus		sp.	H. O'Regan		
11803		Carnivora	Felidae		Panthera		pardus					W. Clyde		
13066		Carnivora	Felidae		Panthera		onca					M. Uhen	4	spe
13066		Carnivora	Felidae		Lynx		rufus		Felis		rufus	M. Uhen	10	spe
13066		Carnivora	Felidae		Miracinonyx	cf.	inexpectatus		Felis	cf.	inexpectata	M. Uhen	2	spe
13066		Carnivora	Felidae		Puma		concolor		Felis		concolor	J. Alroy		
13066		Carnivora	Felidae		Felis		sp.					M. Uhen	1	spe
13293		Carnivora	Felidae		Lynx		sp.					R. Whatley		
13293		Carnivora	Felidae		Panthera		pardus					R. Whatley		
13293		Carnivora	Felidae		Panthera		tigris					R. Whatley		
13456		Carnivora	Felidae		Panthera	cf.	leo					J. Finarelli		
13473		Carnivora	Felidae		Megantereon		praecox					J. Finarelli		
13473		Carnivora	Felidae		Vinayakia		nocturna					J. Finarelli		
13738		Carnivora	Felidae		Panthera		tigris					J. Finarelli	1	spe
13907		Carnivora	Felidae		Leopardus	n. sp.	vorohuensis		Felis	n. sp.	vorohuensis	J. Wertheim	1	ind
17988		Carnivora	Felidae		Pseudaelurus		intrepidus					J. Alroy		
18033		Carnivora	Felidae		Machairodus		coloradensis					J. Alroy		
18034		Carnivora	Felidae		Nimravides	cf.	thinobates					J. Alroy		
18036		Carnivora	Felidae		Machairodus		sp.					J. Alroy		
18046		Carnivora	Felidae		Nimravides		thinobates		Nimravides		catacopis	J. Alroy		
18060		Carnivora	Felidae		Pseudaelurus		sp.					J. Alroy		
18063		Carnivora	Felidae		Pseudaelurus		sp.					J. Alroy		
18072		Carnivora	Felidae		Machairodus		sp.					J. Alroy		
18072		Carnivora	Felidae		Nimravides	cf.	thinobates					J. Alroy		
18073		Carnivora	Felidae		Machairodus	cf.	coloradensis					J. Alroy		
18076		Carnivora	Felidae		Felis		rexroadensis		Felis		rexroadensis	J. Alroy		
18086		Carnivora	Felidae		Nimravides		sp.					J. Alroy		
18097		Carnivora	Felidae		Machairodus	cf.	coloradensis					J. Alroy		
18097		Carnivora	Felidae		Nimravides	n. sp.	hibbardi		Pseudaelurus	n. sp.	hibbardi	J. Alroy		
18097		Carnivora	Felidae		Lynx		proterolyncis		Felis		proterolyncis	J. Alroy		
18126		Carnivora	Felidae		Machairodus		sp.					J. Alroy		
18140		Carnivora	Felidae		Machairodus		coloradensis		Machairodus	cf.	catocopis	J. Alroy		
18140		Carnivora	Felidae		Adelphailuru	n. sp.	kansensis					J. Alroy		
18164		Carnivora	Felidae		Pseudaelurus		intrepidus					J. Alroy		
18165		Carnivora	Felidae		Pseudaelurus		sp.					J. Alroy		
18166		Carnivora	Felidae		Adelphailurus		kansensis					J. Alroy		
18180		Carnivora	Felidae		Nimravides		galiani					J. Alroy		
18181		Carnivora	Felidae		Nimravides		thinobates		Nimravides		catacopis	J. Alroy		
18181		Carnivora	Felidae		Machairodus		sp.					J. Alroy		
18198		Carnivora	Felidae		Nimravides	cf.	thinobates					J. Alroy		
18208		Carnivora	Felidae		Pseudaelurus		sp.					J. Alroy		
18209		Carnivora	Felidae		Pseudaelurus		marshi					J. Alroy		
18219		Carnivora	Felidae		Nimravides		sp.					J. Alroy		
18220		Carnivora	Felidae		Pseudaelurus		stouti		Lynx		stouti	J. Alroy		
18225		Carnivora	Felidae	cf.	Nimravides		sp.					J. Alroy		

Importing data into your new database



Make sure the database you want is selected

Import tab

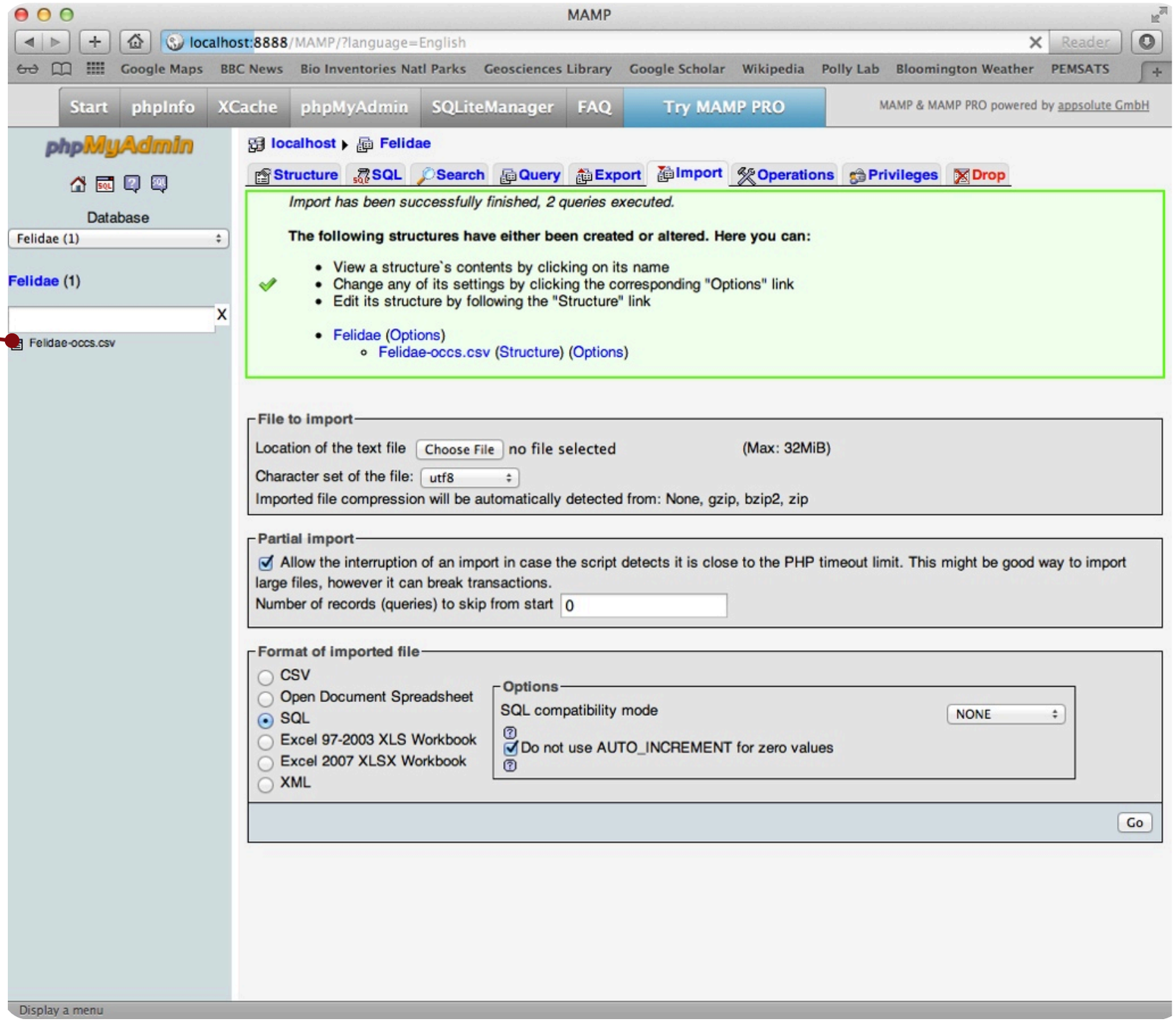
Tick column headers

Select Excel

Start importing

If successful....

New table is listed here



To import data from comma-delimited text file

1. Create new database to be home to your imported tables

2. Select database and click on Import tab

3. Choose file

4. Note that “import” is context dependent. If you have selected database, it will import as new table, but if you have selected a table it will import into that table rather than a new one

5. Once imported, select table from list at left

6. Change table name using the “Operations” tab

7. Add a “key” to your table by selecting your table then selecting the “structure” tab. At bottom add one variable at the beginning of the table. Name that (e.g., “ID”), give it an Integer type (INT), make it the primary index, and tick autoincrement (A_I). This key assigns a unique number to each line in the table, which is important for some operations.

8. Browse data with browse tab. Note that you can delete or edit lines of data using the “X” and the pencil icons.

Browsing and editing data

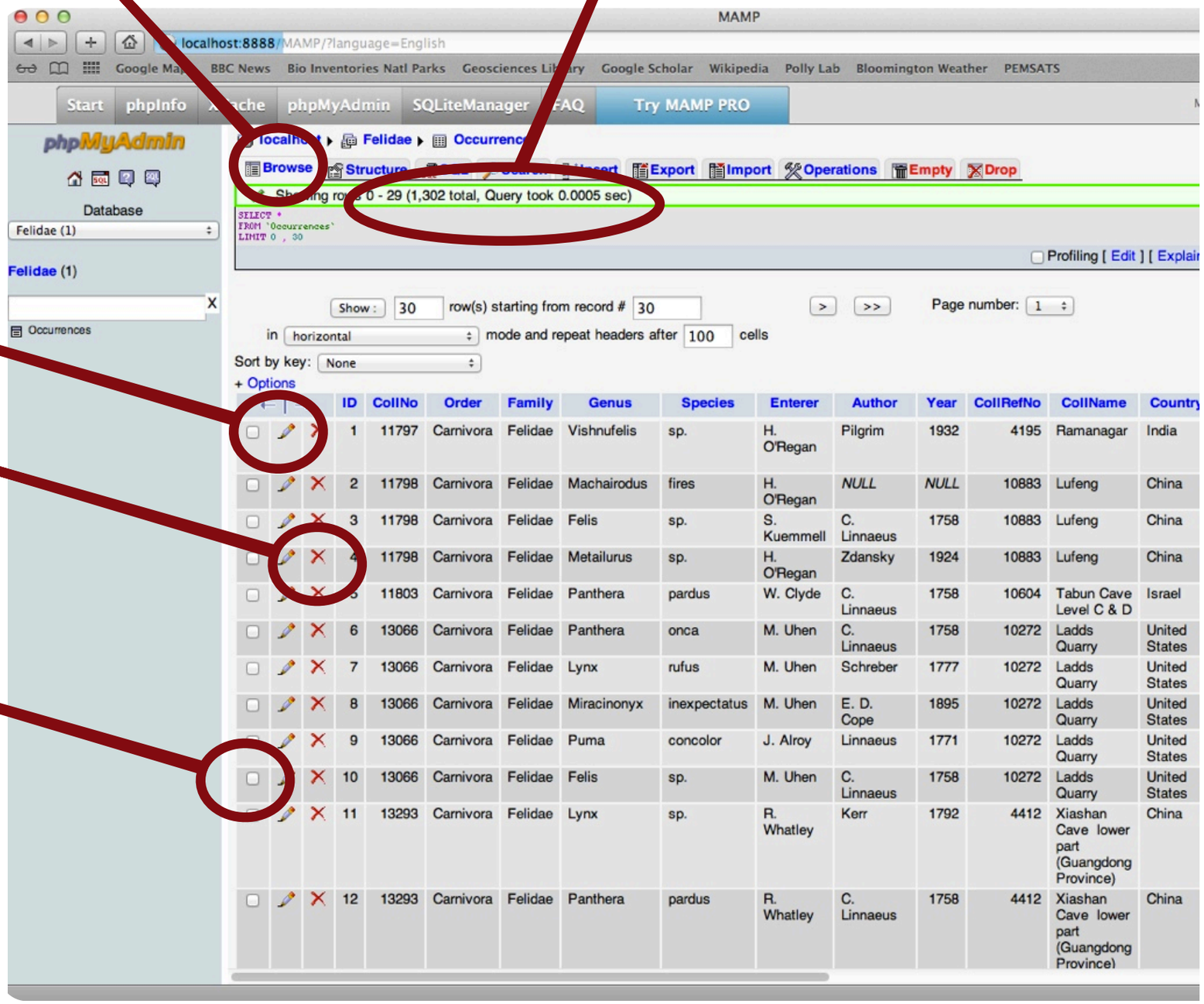
Browse tab

Statistics about data

Click pencil to edit a record

Click x to delete a record

Tick box to select a record for editing or deleting multiple lines with control at bottom of the page

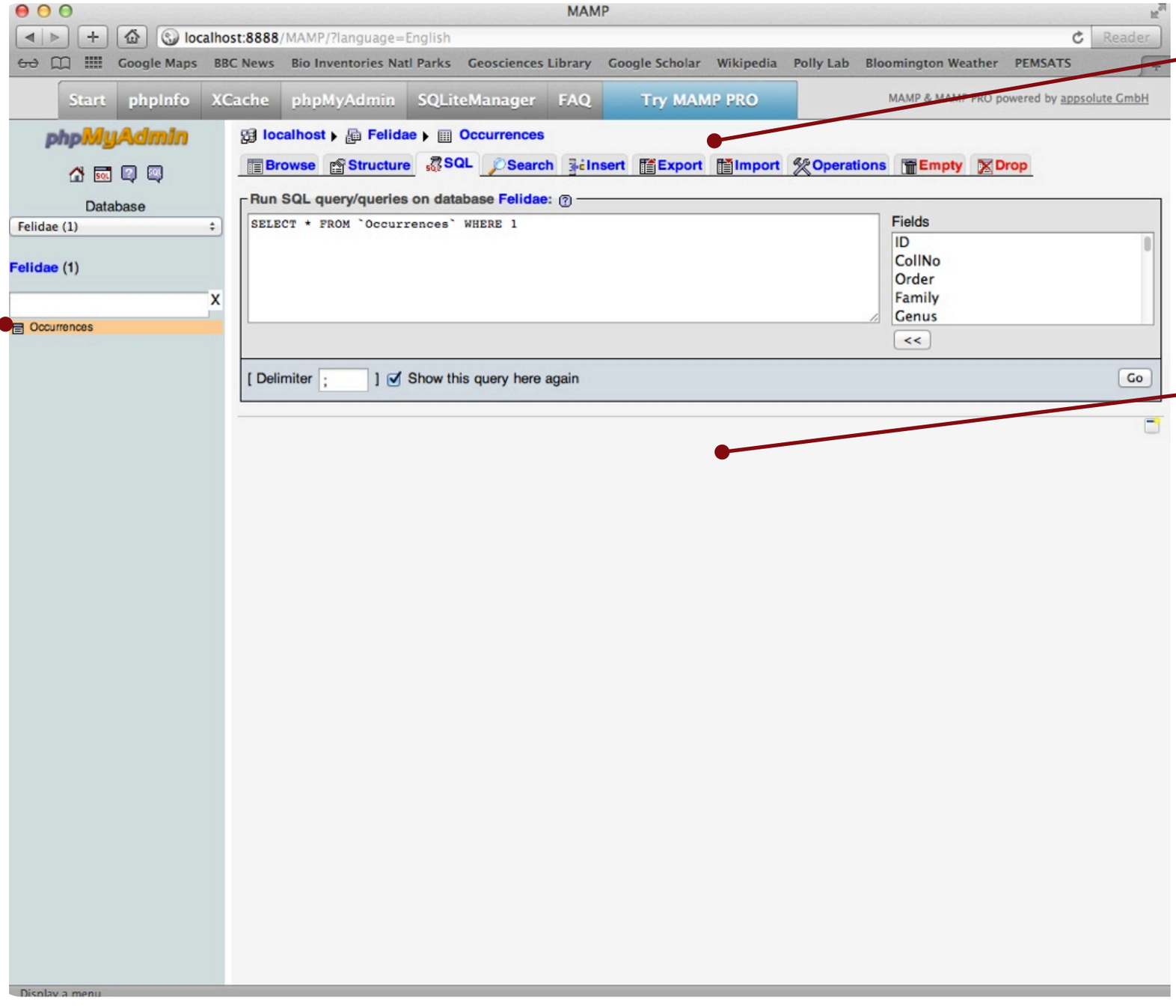


ID	CollNo	Order	Family	Genus	Species	Enterer	Author	Year	CollRefNo	CollName	Country
1	11797	Carnivora	Felidae	Vishnufelis	sp.	H. O'Regan	Pilgrim	1932	4195	Ramanagar	India
2	11798	Carnivora	Felidae	Machairodus	fires	H. O'Regan	*NULL*	*NULL*	10883	Lufeng	China
3	11798	Carnivora	Felidae	Felis	sp.	S. Kuemmell	C. Linnaeus	1758	10883	Lufeng	China
4	11798	Carnivora	Felidae	Metailurus	sp.	H. O'Regan	Zdansky	1924	10883	Lufeng	China
5	11803	Carnivora	Felidae	Panthera	pardus	W. Clyde	C. Linnaeus	1758	10604	Tabun Cave Level C & D	Israel
6	13066	Carnivora	Felidae	Panthera	onca	M. Uhen	C. Linnaeus	1758	10272	Ladds Quarry	United States
7	13066	Carnivora	Felidae	Lynx	rufus	M. Uhen	Schreber	1777	10272	Ladds Quarry	United States
8	13066	Carnivora	Felidae	Miracinonyx	inexpectatus	M. Uhen	E. D. Cope	1895	10272	Ladds Quarry	United States
9	13066	Carnivora	Felidae	Puma	concolor	J. Alroy	Linnaeus	1771	10272	Ladds Quarry	United States
10	13066	Carnivora	Felidae	Felis	sp.	M. Uhen	C. Linnaeus	1758	10272	Ladds Quarry	United States
11	13293	Carnivora	Felidae	Lynx	sp.	R. Whatley	Kerr	1792	4412	Xiashan Cave lower part (Guangdong Province)	China
12	13293	Carnivora	Felidae	Panthera	pardus	R. Whatley	C. Linnaeus	1758	4412	Xiashan Cave lower part (Guangdong Province)	China

Break

Querying your database with SQL

Make sure the database and table you want is selected



Import tab

Tick column headers

Structured Query Language (SQL)

SQL is a standardized way to format instructions to a database server. It resembles English in its grammar and syntax.

Verb (required), Object (required), Object Modifiers (optional),
From Clause (required), Modifying Clause (optional)

```
SELECT  *  FROM  pbdb  WHERE  Genus='Hesperocyon';
```

Select Clause | From Clause | Modifying Clause

Verb gives the command (e.g., “Select”)
Object indicates what fields to select
Clause indicates what table the fields should be selected from
Modifying clause tells what criteria should be used to make the selection

* is the wildcard character, meaning “all fields”

SQL queries

Sentence form:

SELECT *whichever columns you want* **FROM** *whichever table you want* **WHERE** *your criteria are true*

For example, the following query....

SELECT * FROM occurrences

...generates the same browse page you already saw because it selects all columns (*) from the table *taxa* where no criteria are specified (i.e., it selects all rows).

And the following query....

SELECT * FROM occurrences WHERE Country='United States'

selects all the columns from only those rows where the Country column contains the word “United States”.

Finally, the following query....

SELECT Genus FROM taxa WHERE Country='United States'

selects only the taxon name from the rows where the TaxonLevel column contains the word “Species”.

Some powerful modifiers for SQL queries

Select a list of all the unique Countries the database
SELECT DISTINCT Country FROM occurrences

Count the number of entries for each Country
SELECT Country, Count(Country) FROM occurrences GROUP BY Country
note that “group by” is used to define the grouping variable and then the Count() function counts all the entries of Country in each group

Count the number of genera in the database from the USA
SELECT Genus, Count(Genus) FROM occurrences WHERE Country=‘United States’ GROUP BY Genus

Useful functions that can be used in the SELECT statement in conjunction with a GROUP BY statement include: Count(), Avg(), Max(), Min(), Std(), Sum(), Variance(). You can also use multiplication, division, addition, etc. (e.g., “Sum(TaxonID)+Count(TaxonLevel)”).

Quickly tabulate data with SQL

Genus	Count(Genus)
Metailurus	1
Pratifelis	1
Xenosmilus	2
Dinofelis	4
Adelphailurus	5
Megantereon	7
Leopardus	7
Felis	14
Nimravides	19
Miracinonyx	26
Machairodus	26
Pseudaelurus	30
Homotherium	33
Puma	35
Panthera	48
Smilodon	53
Lynx	68

Assignment for next week

Summarize your data

Use SQL and phpMyAdmin to summarize your data. Compile the following information and summarize it neatly, perhaps putting some of the information in a nicely formatted table:

1. Name of your group
2. Total number of occurrences in your database table
3. Number of distinct species
4. Number of distinct genera
5. Number of distinct families
6. Continents represented in your database
7. Number of distinct localities (i.e., distinct latitude and longitude combinations)
8. Using one age variable (e.g., median age), the maximum and minimum age of the occurrences in your data set
9. Number of distinct ages

Your data should have at least a few hundred occurrences, minimum of 5-8 distinct taxa at species or genus level, minimum of 20 distinct localities, and minimum of 10 distinct ages.

Reading (available in OnCourse)

Sepkoski Jr, J. 1993 Ten years in the library: new data confirm paleontological patterns. Paleobiology 19, 43-51.

Lloyd, G., Davis, K., Pisani, D., Tarver, J., Ruta, M., Sakamoto, M., Hone, D., Jennings, R. & Benton, M. 2008 Dinosaurs and the Cretaceous terrestrial revolution. Proceedings of the Royal Society B: Biological Sciences 275, 2483-2490.