Homework Assignment 9

GROUP 4: Bo Haidara, Jonathan Kim, Josh Mikombo

This is a group assignment! Please keep the same partners from HW8 unless you have a good reason to change your group. When you submit this assignment, list the students participating in your group.

Please prepare documentation for your project. The document should be about 5-10 pages long. You can add figures to display the ER diagram and samples of the data that you use. In the document, describe the following:

* Give a description of your database

Our group database links animals from real zoo locations, such as Broussard, Louisiana, Asheboro, North Carolina, and Abilene, Texas, with additional information not previously provided. For each zoo animal, related scientific literature and genetic data can be explored. This can be utilized for medical and educational purposes, offering details on common health risks or the genetic makeup of an animal. The database also connects genetic data supplied by NCBI, revealing some genetic makeup of specific zoo animals. This information is more educational and readily accessible for zoo personnel, researchers, students, and curious individuals.

Our decision to include genetic data was influenced by the global pandemic and the limited public knowledge available on zoonotic viruses that is both accessible and comprehensible for the general public. Often, the information is either too dense, intended only for scholars, or not broad enough, limited to certain animals native to specific areas. The sequencing data of the animals in our database enables researchers and students alike to access valuable information for research purposes. Gaining insight into the genetics of zoo animals is crucial for several reasons, such as understanding future outbreaks or visualizing how certain animals are related, even if they don't appear to be.

Understanding the genetic underpinnings of an organism's traits, its interactions within its natural habitat, and the simulated ecosystem within the zoo allows for a more comprehensive study on all levels. This leads to a better understanding and more accurate information. The spread of false information during the pandemic was largely due to a lack of accessible, reliable data, leaving people's opinions unopposed and unchallenged without proper information.

* Give a rationale for creating the database. Who would use the database and for what purpose? Describe a typical user of your database.

Our database aims to provide the general public with information about animals that are local to their region in America, or those that are closely related to their animal family. Through different forms of educational information such as classification, zoo locations, genetic sequence, and literature references, we hope to educate people about these creatures.

The database contains a wealth of detailed information for each animal, categorized according to three different regions in the United States: Broussard, Louisiana; Asheboro, North Carolina; and Abilene, Texas. These locations simulate different climates and can help people learn about animals in a variety of environments, from coastal cities to mountainous regions.

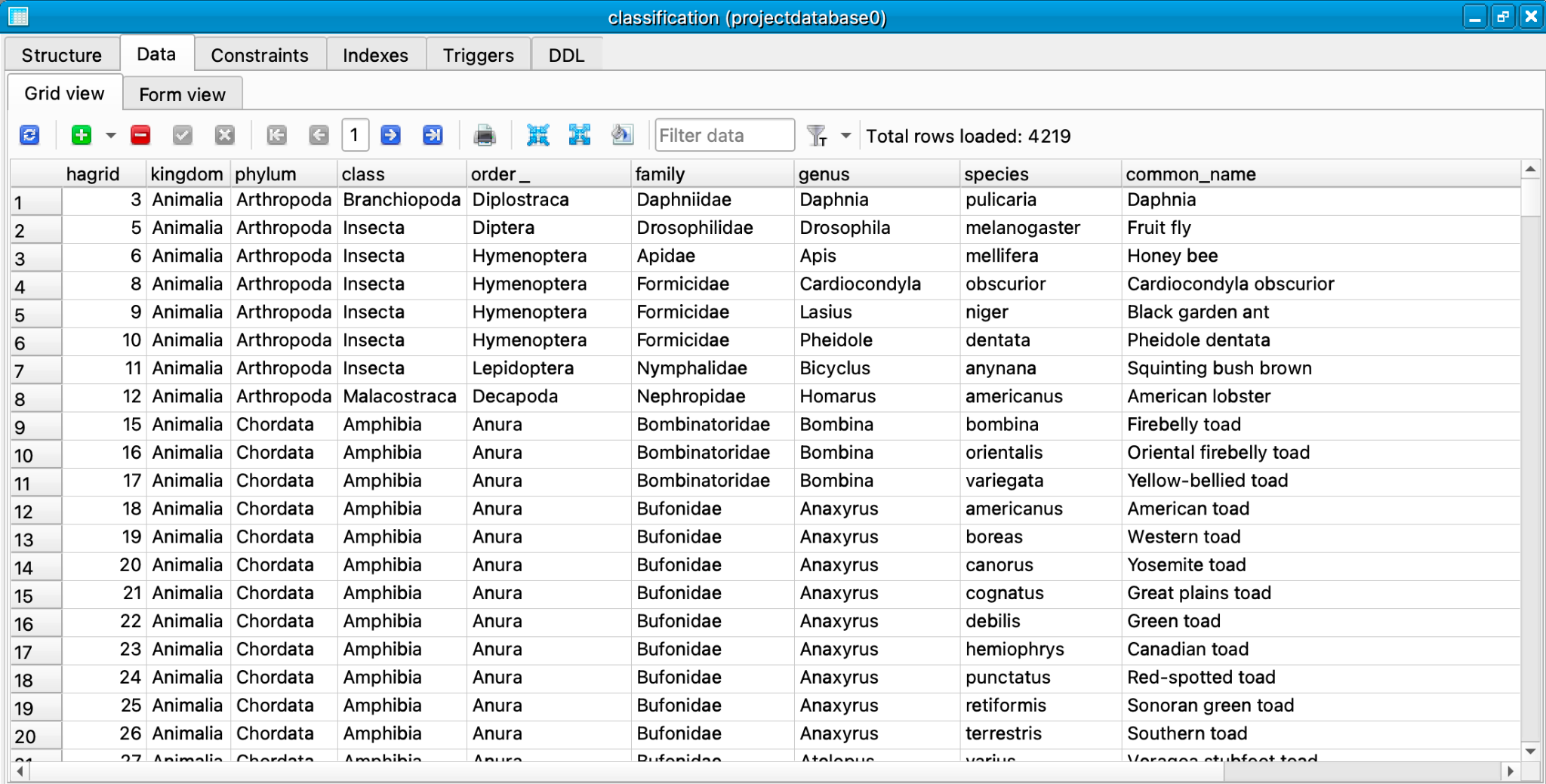
Each zoo in these regions should have animals that are native to the area, as well as exotic species. Our database aims to bridge the gap between resources and zoo animals, providing access to broad and niche information about each creature. Users can access journals about the microscopic details of a particular animal, as well as general information about different species.

The typical user of this database could be a nature conservationist, bird watcher, PETA member, student at different levels of education, or a curious person in the general public who wants to learn more about local animals. For nature conservationists, our database provides information about successful reintroductions of certain animals in the United States, as well as examples of the negative consequences of inbreeding and lack of genetic diversity.

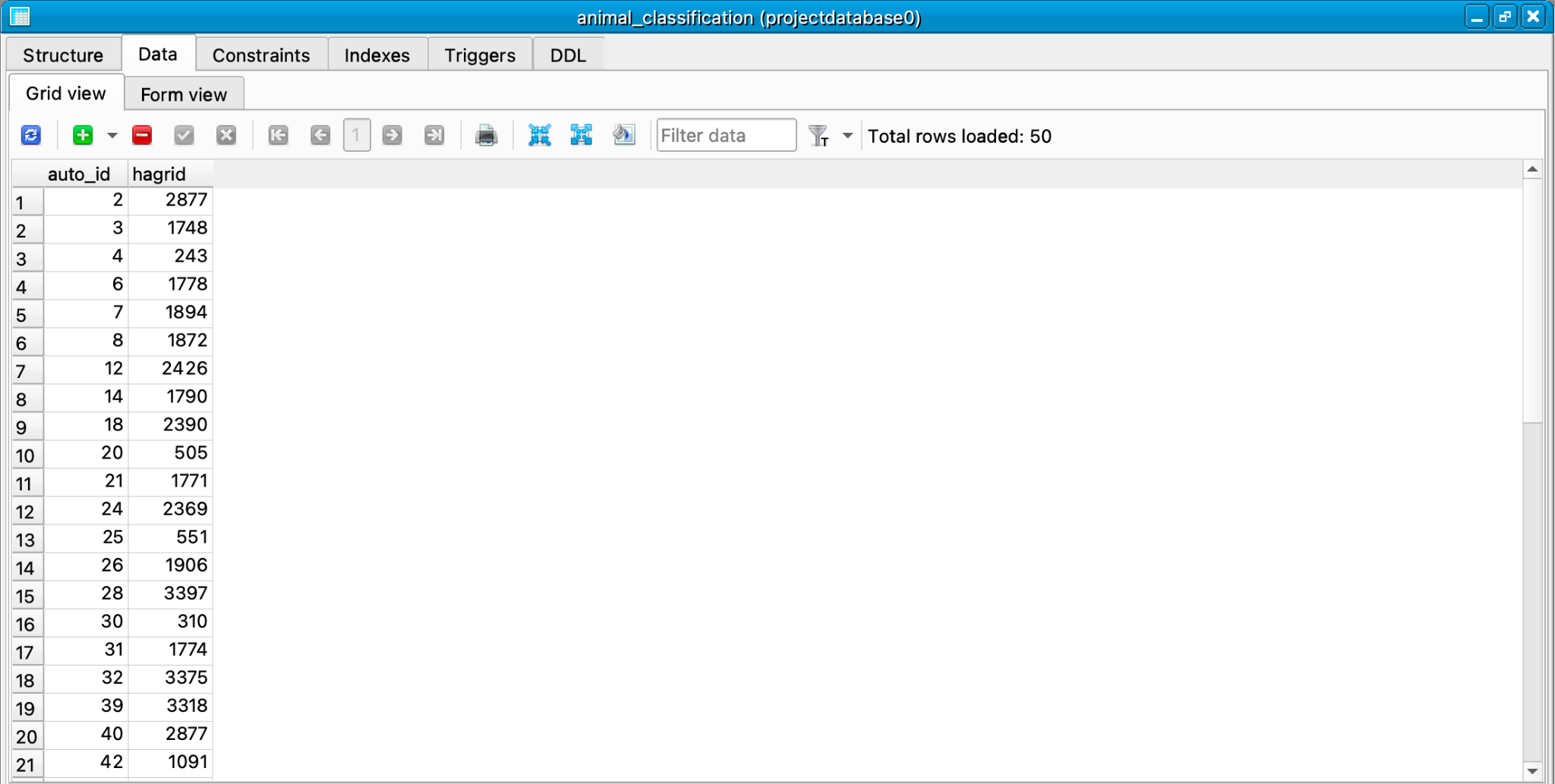
Students at different education levels can also benefit from our database, which provides a more robust understanding of the animal kingdom. Visual learners can see how animals are related and why, while auditory, reading, and writing learners can access literature specific to the animal they are interested in without having to sift through search engines. Kinesthetic learners can navigate the database to find fascinating information about animals in their region.

In conclusion, our database provides a valuable resource for anyone interested in learning more about animals in the United States. By providing detailed information about local species and related animals, we hope to educate people about the animal kingdom and encourage conservation efforts.

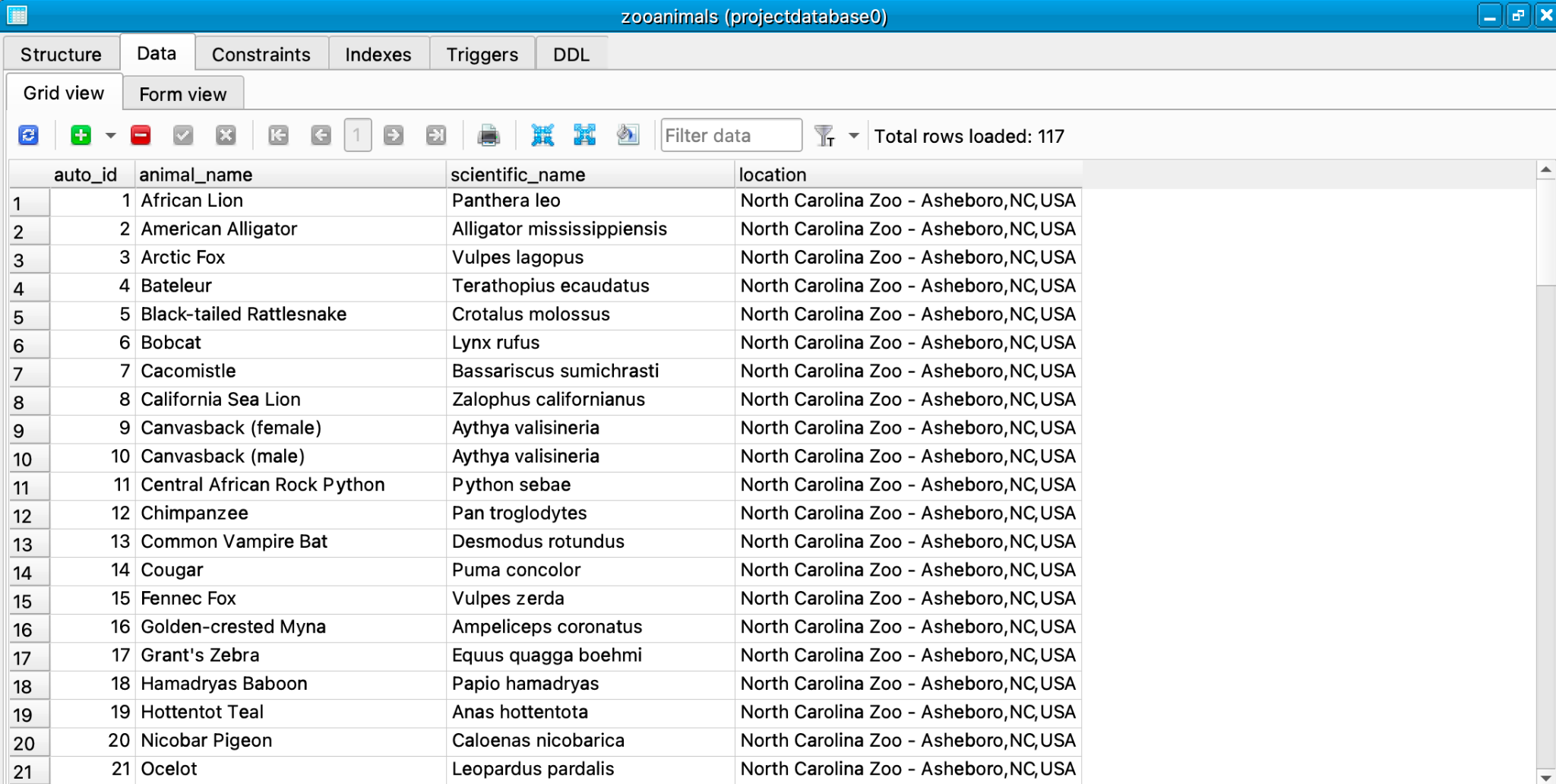
* Show a few examples of the data that will be stored in the database. Explain how you had to modify that data.



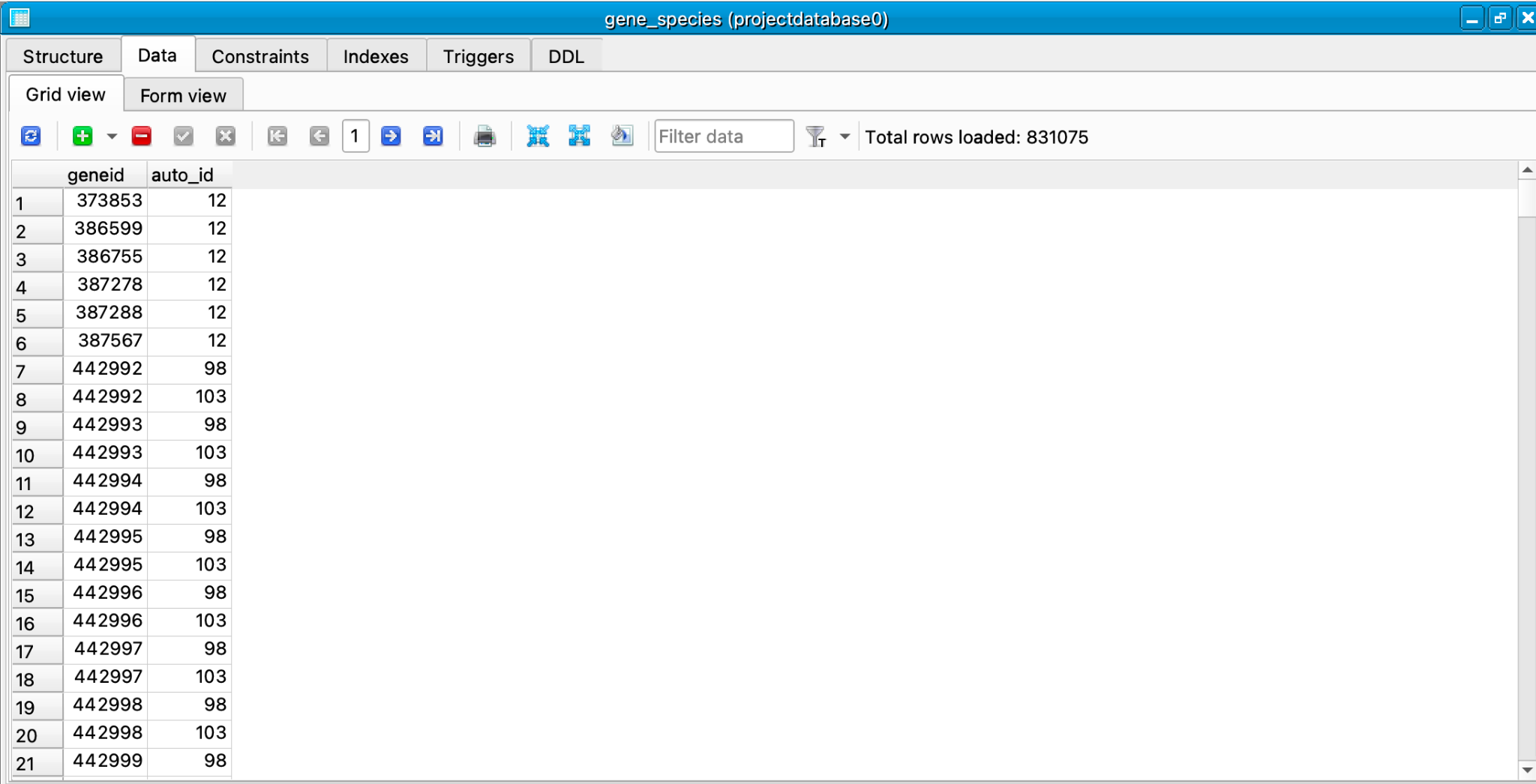
In the classification dataset users can have clear information about the K.P.C.O.F.G.S. and other detailed information about each animal within the database



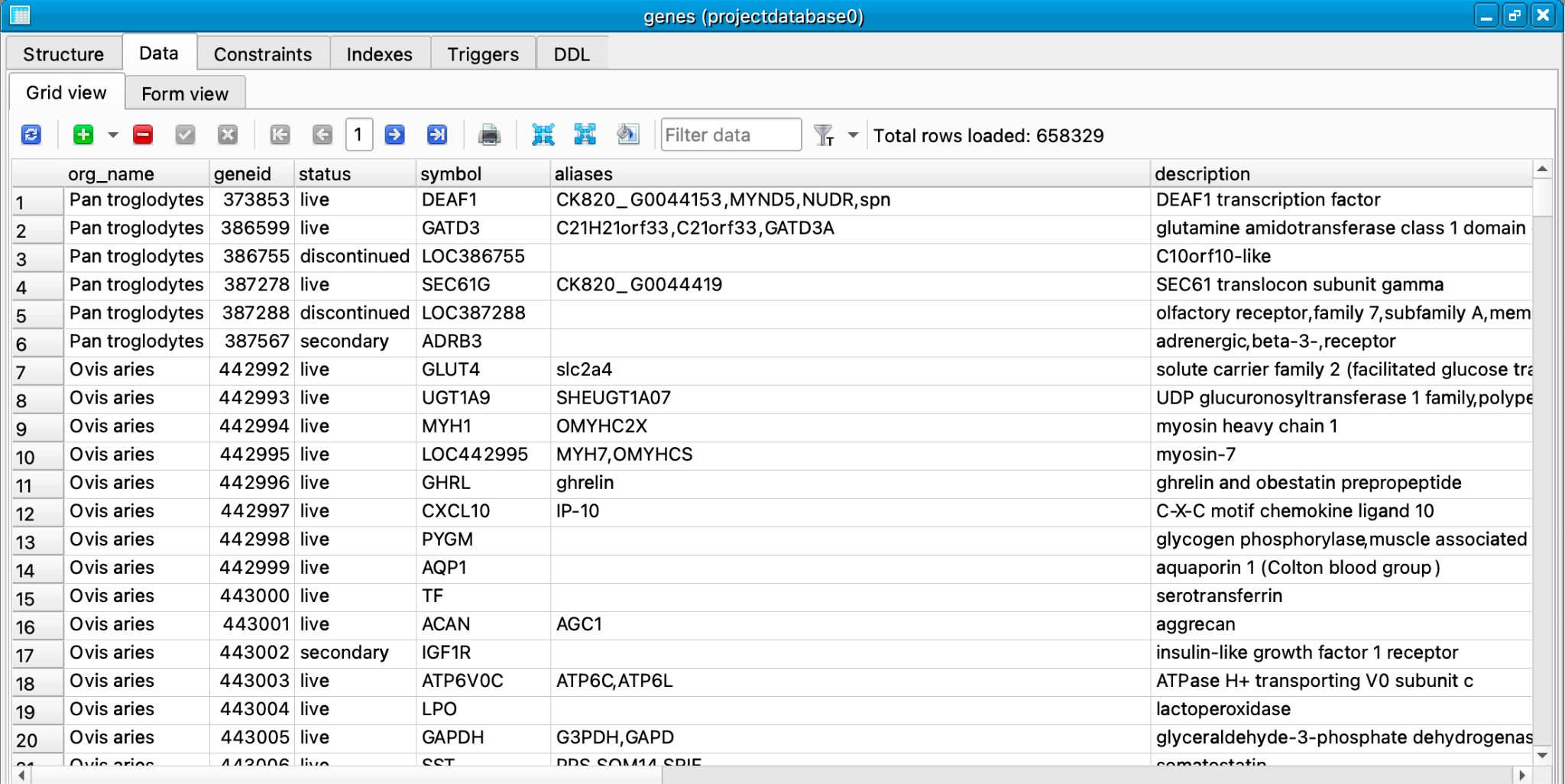
This is the linking table for classification and zooaniamals datasets which will not be visible to the users



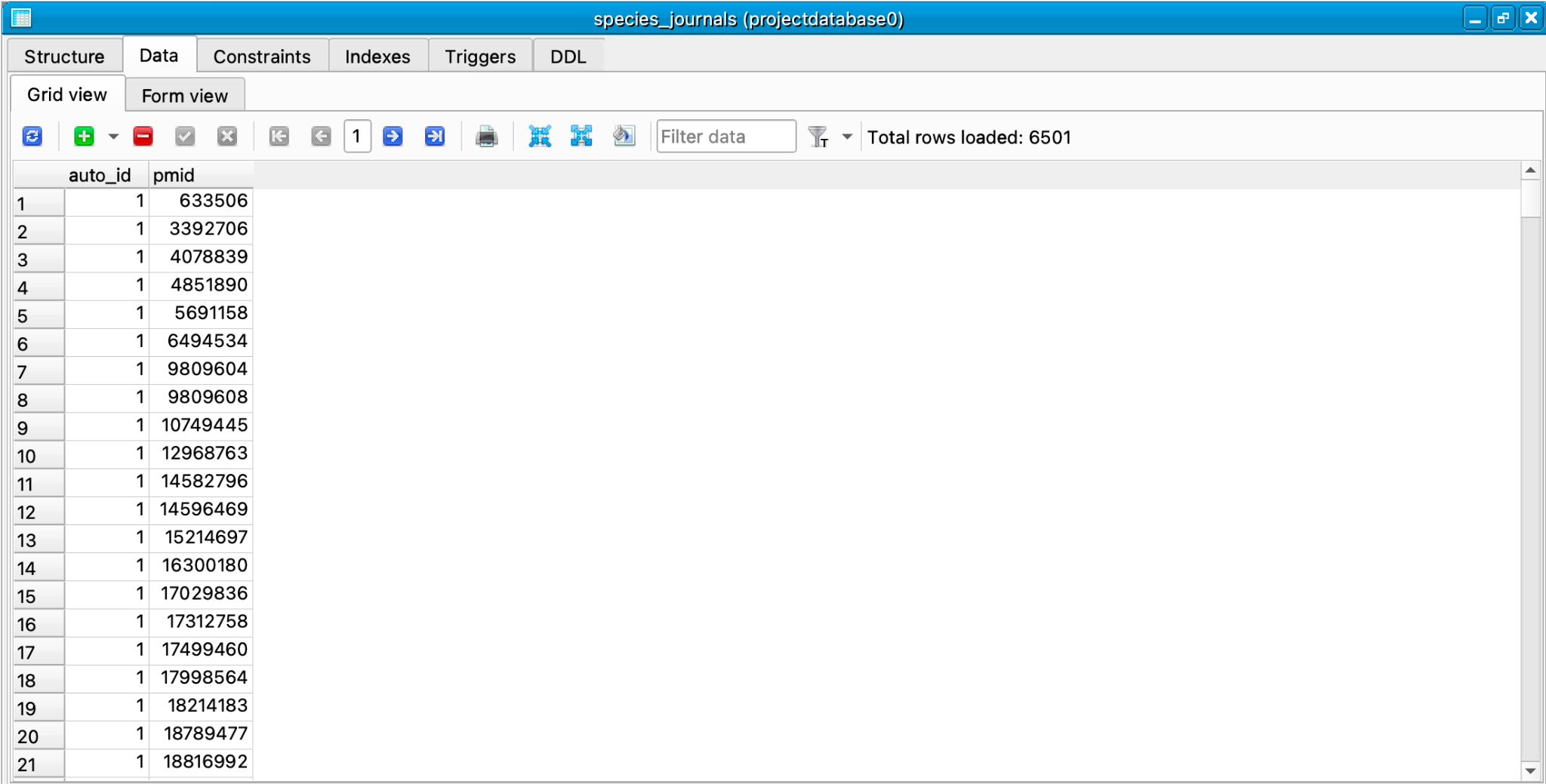
Zooanimals is the main/starting point for each user and is connected to each other data sets (Classification, Genes and Journals) through linking tables



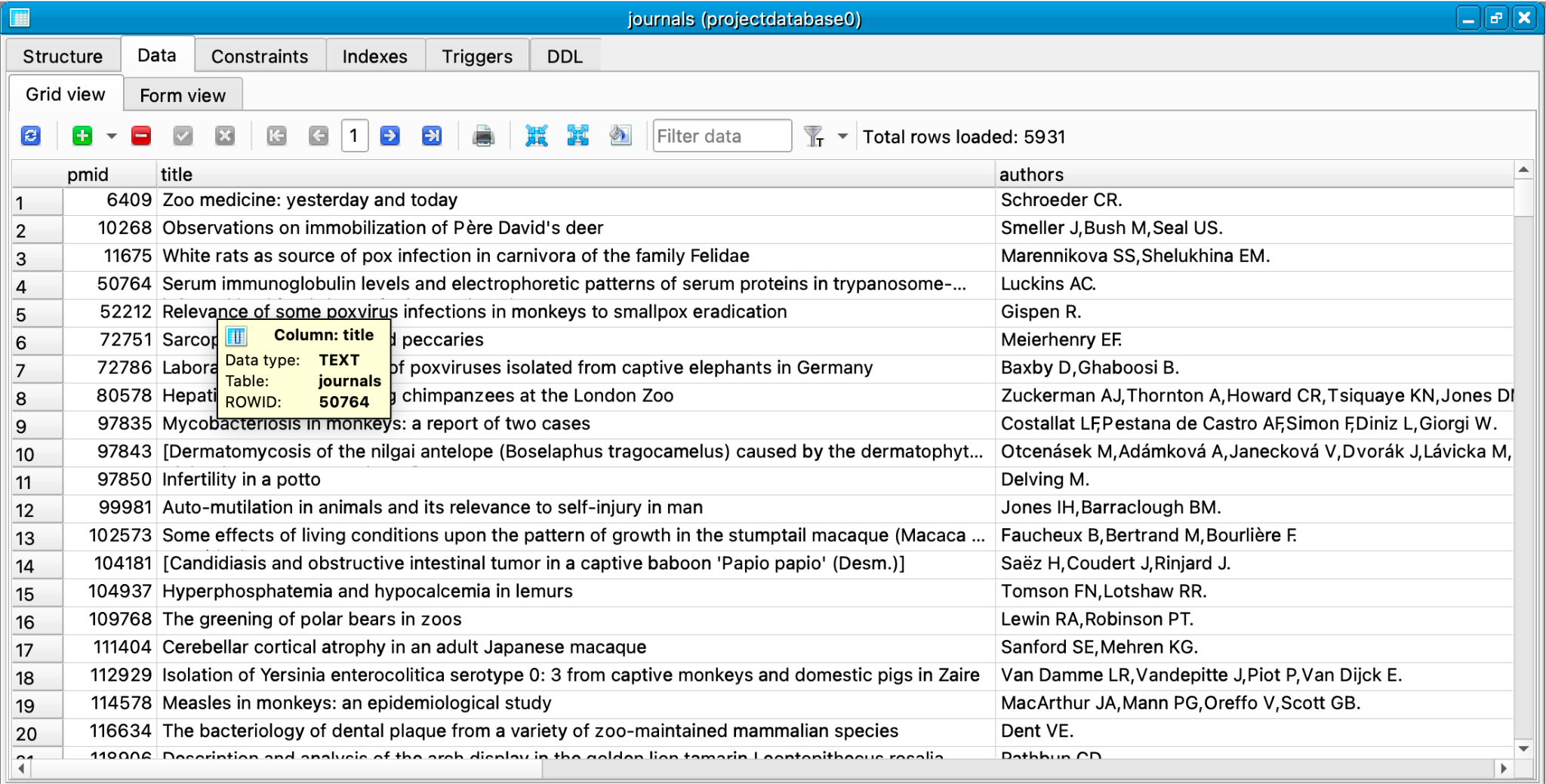
This is the connection table between zooanimals datasets and genes dataset. This is not visible to the users



This is the genetic information about the zooanimals that are in the dataset. This gives the status of the genetic information along with aliases and detailed information about it.

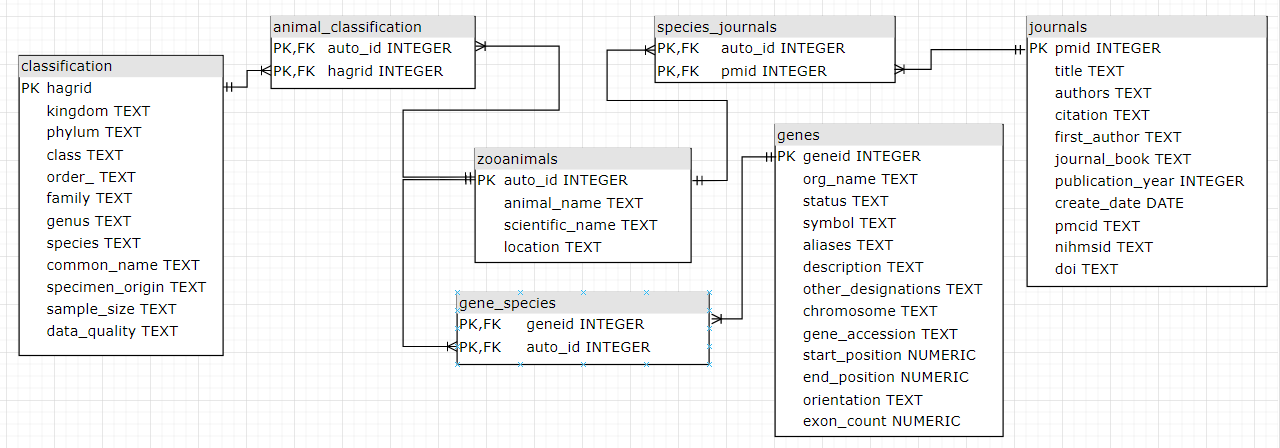


This is a connection table between journals and zooanimals. This will not be visible to the users.



Most of the information and data was cut using R and Python to not include information not useful to the users. This included data that was mostly full of NULL values and information of no importance like temperature and litter size that is not general enough to be consistent with other data across the entire database.

* Plot the Entity-Relationship diagram of the database and explain all the entities and their relationships.



* The zoo animals table is in a one-to-many relationship with animal\_classifcation, species\_journals, and gene\_species.
* The genes table is in a one-to-many relationship with gene\_species.
* The journals table is in a one-to-many relationship with species\_journals.
* The classification table is in a one-to-many relationship with animal\_classification.
* List several typical queries to your database and demonstrate the results of their execution.
  + *Query to search for a lion's scientific name and location(s)*:

**SELECT zooanimals.scientificname, classification.location**

**FROM zooanimals**

**JOIN animal\_classification ON zooanimals.auto\_id = animal\_classification.auto\_id**

**JOIN classification ON animal\_classification.hagrid = classification.hagrid**

**WHERE zooanimals.animal\_name = 'lion' AND classification.scientificname = 'Panthera leo';**

* + *Query to display the zoo with the most animals*

**SELECT location, COUNT(\*) AS animal\_count**

**FROM zooanimals**

**GROUP BY location**

**ORDER BY animal\_count DESC**

**LIMIT 1;**

* + *Query to list the kingdoms found in the North Carolina Zoo in descending order:*

**SELECT classification.kingdom, COUNT(\*) AS animal\_count**

**FROM zooanimals**

**JOIN animal\_classification ON zooanimals.auto\_id = animal\_classification.auto\_id**

**JOIN classification ON animal\_classification.hagrid = classification.hagrid**

**WHERE zooanimals.location = 'North Carolina Zoo'**

**GROUP BY classification.kingdom**

**ORDER BY animal\_count DESC;**

* + *Query to list the journals that published articles on a given species:*

**SELECT journals.title, journals.first\_author, journals.publication\_year**

**FROM journals**

**JOIN species\_journals ON journals.pmid = species\_journals.pmid**

**WHERE species\_journals.scientific\_name = 'species\_name';**

* + *Query to list the scientific names of animals that have genes located on a specific chromosome:*

**SELECT DISTINCT classification.scientificname**

**FROM classification**

**JOIN animal\_classification ON classification.hagrid = animal\_classification.hagrid**

**JOIN zooanimals ON animal\_classification.auto\_id = zooanimals.auto\_id**

**JOIN genes ON genes.org\_name = zooanimals.animal\_name**

**WHERE genes.chromosome = 'chromosome\_name';**

* + *Query to list the common names of animals that are located in a given zoo and belong to a given family:*

**SELECT DISTINCT classification.common\_name**

**FROM classification**

**JOIN animal\_classification ON classification.hagrid = animal\_classification.hagrid**

**JOIN zooanimals ON animal\_classification.auto\_id = zooanimals.auto\_id**

**WHERE zooanimals.location = 'zoo\_name' AND classification.family = 'family\_name';**

* + *Query to list the genes that have a certain keyword in their description:*

**SELECT geneid, symbol, description**

**FROM genes**

**WHERE description LIKE '%keyword%';**

* + *Query to list the number of articles published each year for a given journal:*

**SELECT publication\_year, COUNT(\*) AS article\_count**

**FROM journals**

**WHERE journal\_book = 'journal\_name'**

**GROUP BY publication\_year;**

* + *Query to list the genes that are associated with a certain disease:*

**SELECT genes.symbol, genes.description, genes.other\_designations**

**FROM genes**

**JOIN gene\_species ON genes.geneid;**

* + *Query to breakdown the number of animals in each species in the database in descending order:*

**SELECT classification.species, COUNT(\*) AS num\_animals**

**FROM animal\_classification**

**JOIN classification ON animal\_classification.hagrid = classification.hagrid**

**GROUP BY classification.species**

**ORDER BY num\_animals DESC;**

* Describe what difficulties/issues you had while working on this project, and how you resolved them.
  + First and foremost, finding relevant datasets and information to populate the database.
    - This was solved by finding websites with database-readable and not database-readable information in a format that could be translated later.
  + Translating certain files like XML and HTML into data-readable files like CSV to be imported specifically into SQLiteStudio.
  + Selecting only certain columns from data to include data pertinent to our database rather than all columns that may not be useful for us and our users.
    - Used conjunction of R, Python, and SQL to solve both problems above.
  + Figuring out how to connect these separate datasets and ideas of literature, genes, and journals to the animals from a particular zoo.
    - Solved through creating a connecting/linking tables in between.

**Links/References**

* Broussard, Louisiana (Zoosiana)
  + Zoosiana. (n.d.). Home. Retrieved from https://zoosiana.com/.
* , Asheboro, North Carolina (North Carolina Zoo)
  + North Carolina Zoo. (n.d.). Home. Retrieved from https://www.nczoo.org/.
* Abilene, Texas (Abilene Zoo)
  + Abilene Zoological Gardens. (n.d.). Home. Retrieved from https://abilenezoo.org/.
* ER Diagram Maker
  + diagrams.net. (2022, February 3). How to insert SQL into your diagrams. Retrieved from <https://www.diagrams.net/blog/insert-sql>.
* NCBI GENES RESULT
  + National Center for Biotechnology Information. (n.d.). Gene - NCBI. Retrieved April 4, 2023, from <https://www.ncbi.nlm.nih.gov/gene?term=%22Equus%20quagga%20boehmi%22%20%5BOrganism%5D%20OR%20%22Oxyura%20jamaicensis%22%20%5BOrganism%5D%20OR%20%22Ampeliceps%20coronatus%22%20%5BOrganism%5D%20OR%20%22Naja%20pallida%22%20%5BOrganism%5D%20OR%20%22Basiliscus%20plumifrons%22%20%5BOrganism%5D%20OR%20%22Aythya%20valisineria%22%20%5BOrganism%5D%20OR%20%22Dendrocygna%20bicolor%22%20%5BOrganism%5D%20OR%20%22Macrochelys%20temminckii%22%20%5BOrganism%5D%20OR%20%22Zalophus%20californianus%22%20%5BOrganism%5D%20OR%20%22Anser%20anser%22%20%5BOrganism%5D%20OR%20%22Erythrocebus%20patas%22%20%5BOrganism%5D%20OR%20%22Puma%20concolor%22%20%5BOrganism%5D%20OR%20%22Struthio%20camelus%22%20%5BOrganism%5D%20OR%20%22Phacochoerus%20sp.%22%20%5BOrganism%5D%20OR%20%22Leopardus%20pardalis%22%20%5BOrganism%5D%20OR%20%22Giraffa%20camelopardalis%20reticulata%22%20%5BOrganism%5D%20OR%20%22Trioceros%20melleri%22%20%5BOrganism%5D%20OR%20%22Leptailurus%20serval%22%20%5BOrganism%5D%20OR%20%22Coendou%20prehensilis%22%20%5BOrganism%5D%20OR%20%22Graptemys%20versa%22%20%5BOrganism%5D%20OR%20%22Lachesis%20stenophrys%22%20%5BOrganism%5D%20OR%20%22Bassariscus%20sumichrasti%22%20%5BOrganism%5D%20OR%20%22Python%20regius%22%20%5BOrganism%5D%20OR%20%22Leiothrix%20lutea%22%20%5BOrganism%5D%20OR%20%22Mycteria%20ibis%22%20%5BOrganism%5D%20OR%20%22Amazona%20finschi%22%20%5BOrganism%5D%20OR%20%22Panthera%20leo%22%20%>
* Animal aging and life history website
  + de Magalhães, J. P., & Costa, J. (2021). Annotated lists of mammals, birds, and fishes exhibiting extreme life spans. Retrieved April 4, 2023, from <https://genomics.senescence.info/species/index.html>
* Animal life span information
  + de Magalhães, J. P., & Costa, J. (2021). Dataset of mammalian, avian and piscine maximum reported lifespan versus age at sexual maturity. Retrieved April 4, 2023, from <https://genomics.senescence.info/species/dataset.zip>
* More zoo animal databases
  + Wild Welfare. (n.d.). Zoo animal databases & associations. Retrieved April 4, 2023, from <https://wildwelfare.org/zoo-animal-databases-associations/>
* Animal article/journal references
  + EBSCOhost. (n.d.). Advanced search - Academic Search Complete. Retrieved April 4, 2023, from <https://search.ebscohost.com/login.aspx?direct=true&AuthType=shib&db=fzh&bquery=animal&authtype=shib&type=1&searchMode=Standard&site=ehost-live&scope=site&custid=s5822979>
* Data scrape from HTML of online zoo website:
  + The Online Zoo. (n.d.). North Carolina Zoo. Retrieved April 4, 2023, from <http://theonlinezoo.com/pages/north_carolina_zoo.html>
  + The Online Zoo. (n.d.). North Carolina Zoo - Zoo Links. Retrieved April 4, 2023, from <http://theonlinezoo.com/pages/north_carolina_zoo.html>
* NCBI journal results of animals from a zoo:
  + PubMed. (n.d.). Results for "'Animals, Zoo'[Mesh]". Retrieved April 4, 2023, from https://pubmed.ncbi.nlm.nih.gov/?term=%22Animals%2C+Zoo%22%5BMesh%5D&size=200