Data Warehousing Final Project - Occurrence Data

Eric Bataller Thomas, Maxim Fedotov, Danilo Méndez Rubio

December 8, 2021

Abstract

In this project we work with data on animal sightings in their natural habitat and want to connect it with data on climate and some extra information describing these animals. The goal is to to have an insight on a state of ecosystem health from this data. For this purpose, we create a datawarehouse on sightings of all carnivore mammals in Spain to show that the concept is scalable and worth consideration as well as to establish directions for further extensions. The project is able to create favorable conditions for interactions of researchers in biology, ecology and related fields. We inferred that the data on animal observations is most likely prepared by professionals which gives us a good quality source of information. We found that there are not many unique *carnivori* observed in Spain, but they sufficiently differ in their traits which might be good sign for biodiversity. Nevertheless, there are interesting insights that could be explored beyond the scope of this project, so we encourage other researchers to contribute into moving science in this direction.

1 Introduction

This project is centered in understanding the wealth of biodiversity of a region, independent of scale, through sightings from different sources: citizen scientists, research institutions, public records and the private sector. For this particular project we will do a first draft to have a proof of concept and an MVP. We will concentrate only on the sightings of animals of order *carnivora* in Spain from the last 21 years (2000-2021).

The Order *carnivora* is comprised of all carnivore mammals. Carnivore mammals, especially large species, are considered as *Keystone* species [2]. This means that their presence has a huge impact on biodiversity and the health of ecological systems[1]. Humans and invasive species have a negative impact in their environment but native carnivores are very important. This is why we are interested in their sightings as a proxy of ecosystem health and to understand the impact of human centers of population have on *carnivora* sightings. This leads us quite naturally to postulate 3 requirements to understand ecosystem health near human settlements:

- Requirement 1 - We want to understand where the animals are sighted and if there is a strong relationship to climate data or if it is more linked to geographical data
- Requirement 2 - We want to understand the strata of species traits in a shared environment, meaning that we want to see where there are many species of *carnivora* with different traits since diversity is an indicator of abundance. The different traits will correspond to different evolutionary specializations, diet and the Status of endangerment of those species.
- Requirement 3 - We are worried about the origin and integrity of the data since the quality of the observations depends on who is sighting the animal. An expert that dedicates their life to the study of such species is a better source of information than a hobby hiker when the time comes to identify the species observed. We want to know when (time of day and day of week) the data is uploaded to understand the quality of our data given the hypothesis that a workday upload during working hours corresponds to an expert and a weekend upload corresponds to citizen scientists

2 Data Sources

2.1 Occurrence Data: *https://www.gbif.org* - Source 1

The Global Biodiversity Information Facility houses occurrence data from a curated selection from a host of different databases into one centralized data warehouse. There are two ways of accessing this information:

1. From *https://www.gbif.org/occurrence/search?country=ES&taxon_key=732&year=2000,2021* where we can filter occurrences through the URL or the website functionality
2. From the *rgbif* R package, which directly connects to GBIF's API and lets the user select and filter a greater variety of variables and Rows. Since the API is accessed throguh R, the format of the data is a "List" object (list of list), which constitutes the equivalent .json file format available in other lenguages.

We chose to obtain the data through the API, were we chose the following variables from the over 100 available options: key, scientificName, decimalLatitude, decimalLongitude, basisOfRecord, occurrenceStatus, kingdom, phylum, class, order, family, genus, species, datasetName, rightsHolder, genericName, specificEpithet, dateIdentified, stateProvince (this is the raw data, and some of these fields will be dumped trhougout the ETL workflow). For our project we selected all the carnivores that have been observed in Spain from 2000 to 2021. We had 40445 rows/observations. In ulterior sections we will explain how we used every variable.

2.2 Town Data: *https://www.businessintelligence.info/varios/longitud-latitud-pueblos-espana.html* - Source 2

longitud-latitud-pueblos-espana.html gathers data on the location and altitude of ALL towns in Spain. The town data was downloaded via direct link and was provided in an XML format. The columns of this data set are: Comunidad, Provincia, Población, Latitud, Longitud, Altitud, Habitantes, Hombres and Mujeres.

We aim to infer the climate data of each occurrence based on the climatic conditions of its closest town. This allows us to be more efficient in terms of storing the climatic conditions as we don't have to check the climate conditions for the coordinates of each occurrence but rather for all towns in spain which covers most of the surface (as it wouldn't make sense given that there is a specific resolution on the climatic data - as mentioned in the next subsection).

We are not interested in the proportion of males and females in the towns so we will not use this information in our analysis.

2.3 Climate Data: *https://worldclim.org/* - Source 3

WorldClim is a database of high spatial resolution global weather and climate data. We used this data to obtain historical weather conditions in the coordinates of all towns in Spain. Users can access this data by downloading it directly from the website (*https://worldclim.org/data/worldclim21.html*) or through the API that can be accessed with *raster* R package. The output of the API is a "List" object in R.

Using this API we can access the Bioclimatic data: Annual Mean Temperature, Mean Diurnal Range, Isothermality, Temperature Seasonality, Max Temperature of Warmest Month, Min Temperature of Coldest Month, Temperature Annual Range, Mean Temperature of Wettest Quarter, Mean Temperature of Driest Quarter, Mean Temperature of Warmest Quarter, Mean Temperature of Coldest Quarter, Annual Precipitation, Precipitation of Wettest Month, Precipitation of Driest Month, Precipitation Seasonality, Precipitation of Wettest Quarter, Precipitation of Driest Quarter, Precipitation of Warmest Quarter and Precipitation of Coldest Quarter. These are standard variables for climate analysis.

We can also access elevation data from this source, we just get the meters over sea level. Both of these sources have a spatial resolution, meaning that they average the variables over an area, the smaller the resolution the smaller the area. We chose a 5 minute arc which corresponds to a 7.1km by 7.1km (50 km square) at Spain's latitude.

2.4 Mammal Data : *https://megapast2future.github.io/PHYLACINE_1.2/* - Source 4 and 5

We wish to understand more about the animals we are analysing so we wish to add more information. We realised that some animals have more than one scientific name depending on the taxon classification given as there is some debate on the correct nomenclature of certain species. Fortunately for us, **PHYLACINE** provides both detailed data about the animal traits of ALL mammals, along with a separate dataframe containing major synonyms of each species. We downloaded *Synonyms* (source 4) for these species so that we can identify the species name each occurrance refers to and we can set them all to the same nomenclature. We downloaded the synonym data directly from the web page as a csv.

Next we downloaded a *traits* (source 5) data set directly from the web page as a csv. This data set has the following variables: Binomial, Order, Family, Genus, Species, Terrestrial, Marine, Freshwater, Aerial, Life.Habit.Method, Mass, Mass.Method, Mass.Comparison, Island.Endemicity, IUCN.Status, Diet.Plant, Diet.Invertebrate and Diet.Vertebrate. This data will help us understand the type of mammal we are studying and it will let us drill down on their characteristics.

3 Entity-Relationship Model

In this section we discuss Entity-Relationship model corresponding to our database. The diagram representing ER model is shown on Figure 1. In our setting we have 5 entities, namely: occurrences (animal sightings), animals, towns, observers, datasets. These entities relate to each other via the following relationships. Observer records an occurrence into a database. Each occurrence has an associated town, which is the closest one in sense of Haversine distance. Each occurrence is per se a sighting of a specific animal, that is – it identifies an animal.

It is possible that for a specific occurrence there is either no dataset or no observer specified, that is why connection between occurrences and observers is many (mandatory) to one (optional) as well as connection between occurrences and datasets. The ordinality ensures that there is at least one animal sighting connected with existing datasets and observers, but a dataset or observer can be missing. Since occurrence identifies an observation of a specific animal, it is connected as many (optional) to one: an occurrence always reveals an existing animal, but it might happen that some animals are not observed. The same logic applies for connection between occurrences an towns: each occurrence is associated with a specific time by the closest Haversine distance, but for some towns we might not have associated observations at all.

The attributes are grouped in broader categories just to make the picture cleaner and convey the content in a more transparent way. The whole range of attributes is specified it Star schema. Nevertheless, we also display them in this section to show which attributes correspond to the attribute groups shown in the ER-model. In the following explanation, we keep the names of attributes to be the same as they are indicated in our database.

The **occurrences** are the main entities of our interest. Each occurrence has the following groups of attributes:

- *Animal appellation*
  It consists of order, family, genus, and species names.
- *Date and time information*
  This group specifies a date, an hour, and a weekday when the occurrence was recorded in a database.
- *Location*
  A pair of coordinates: a longitude and a latitude.

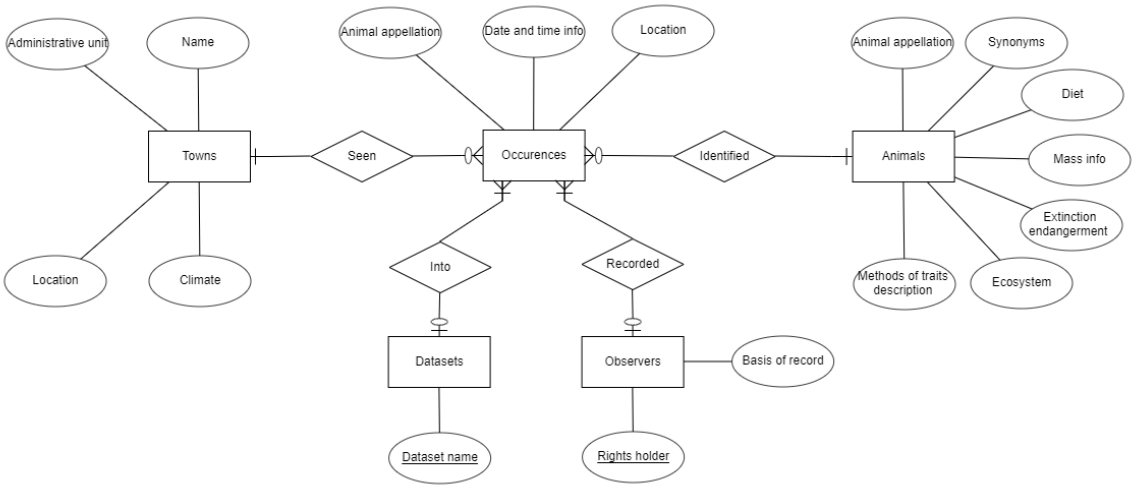

Figure 1: ER model

As we mentioned before, an occurrence identifies a specific animal. An **animal** has the following attributes:

- *Animal appellation*
  The same as in occurrence, but in addition contains also binomial of an animal.
- *Synonyms*
  Contains synonyms of animal appellations because some animals may be called in different ways. So, most of the major databases also contain different appellations for animals. Namely, in our case these attributes are Genus 1.1, Species 1.1, Genus 1.0, Species 1.0, EltonTraits.1.0.Genus, EltonTraits.1.0. Species, IUCN.2016.3.Genus, IUCN.2016.3.Species.
- *Diet*
  This group contains attributes describing diet of an animal. They specify a percentage of a specific source of food in an animal diet. The attributes are: Diet.Plant, Diet.Vertebrate, Diet.Invertebrate.
- *Mass info*
  Contains attribute Mass.g which specifies mass in grams.
- *Extinction endangerment*
  Has attribute IUCN.status which is an identification of an animal extinction or its risk.
- *Ecosystems*
  Consists of attributes describing a natural habitat of an animal: terrestrial, marine, freshwater, aerial. It also contains Island.Endemicity attribute which indicates whether an animal belongs to a specific Islandic territory.
- *Methods of traits description*
  Attribute Life.Habit.Method specifies how information on a natural habitat of an animal was inferred. Mass.Method tells a way in which a mass data was retrieved. There is also Diet.Method included in this category; it specifies how the information on a diet was obtained.

Towns possess the following groups of attributes:

- *Name*
  Defines a name of a town: población.
- *Administrative unit*
  Contains attributes comunidad and provincial.

- *Location*
  Latitude, longitude, and height_above_sea attributes.

- *Climate*
  All the climatic data is unified by this group. It contains attributes which specify temperature, isothermality, precipitation rates, and diurnal hours which correspond to each town. Namely, the attributes are: annual_mean_temperature, mean_diurnal_range, isohermality, temperature_seasonality , max_temperature_of_warmest_month, min_temperature_of_coldes_month, temperature_annual_range, mean_temperature_of_wettest_quarter, mean_temperature_of_driest_quarter, mean_temperature_of_warmest_quarter, mean_temperature_of_coldest_quarter, annual_precipitation, precipitation_of_wettest_month, precipitation_of_driest_month, precipitation_seasonality, precipitation_of_wettest_quarter, precipitation_of_driest_quarter, precipitation_of_warmest_quarter, precipitation_of_coldest_quarter.

The **Observer** entity has a pair of attributes:

- *Rights holder*, which is unique because identifies a login of a person who holds rights on an observation.

- *Basis of record*

The **Dataset** entity is described by only one attribute – Dataset name.

It is worth mentioning that, for example, observers are connected with datasets through occurrences since they post information there.

4 Star Schema

In this section we show the Star Schema corresponds to our database as well as shortly discuss the tables and connections presented on it. The schema is displayed on Figure 2. It consists of a fact table, and several dimension tables which the fact one refers to.

The fact table "fact_table_occ" which corresponds to occurrences data in which we are interested the most. The dimension tables are Towns (dim_towns_climate), Animals(dim_animal), Observers (dim_observer), and Datasets (dim_dataset). The tables show all the fields they contain, as well as primary (golden color) and foreign (grey color) keys. We have foreign keys only in the fact (occurrence) table because it refers to the dimension tables. These referring connections are shown via straight black lines. These lines point directly at fields on which the tables are connected. Note that the fields presented correspond to the groups of attributes shown above in the ER model section.

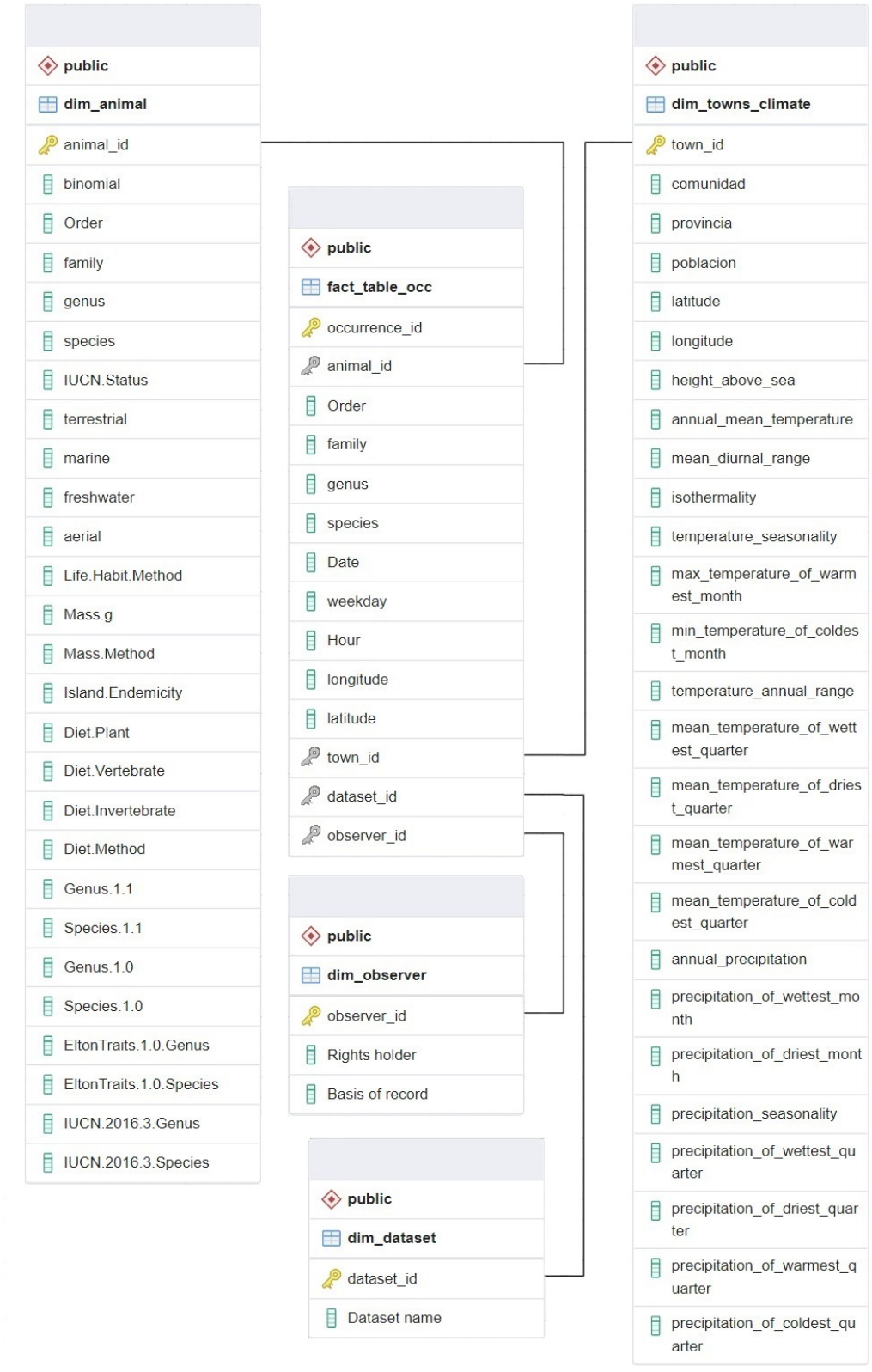

Figure 2: Star Schema

5 Logical Data Map

In this section, describe your logical data map, i.e. how every row of every data source is handled throughout the transformations such that it is a part of your star schema.
Note: The Transformation column in the table above doesn't contain generic transformations used across fields such as transforming 'NA' type Null into actual 'Null', or fixing the length of each field for the sake of standardization (e.g. changing an arbitrary VARCHAR(32) to VARCHAR(35)).

Table 1:

Target				Source		Transformation
Table type	Destination	Column	Data type	Source	Column	
Dimension	dim_animal	Binomial	VARCHAR(35)	5	Binomial.1.2	Retrieved the data without any transformation.
Dimension	dim_animal	Order	VARCHAR(35)	5	Order.1.2	Retrieved the data without any transformation.
Dimension	dim_animal	Family	VARCHAR(35)	5	Family.1.2	Retrieved the data without any transformation.
Dimension	dim_animal	Genus	VARCHAR(35)	5	Genus.1.2	Retrieved the data without any transformation.
Dimension	dim_animal	Species	VARCHAR(35)	5	Species.1.2	Retrieved the data without any transformation.
Dimension	dim_animal	IUCN.Status	VARCHAR(2)	5	=	Retrieved the data without any transformation.
Dimension	dim_animal	Terrestrial	BOOLEAN	5	=	Transformed from int to Boolean.
Dimension	dim_animal	Marine	BOOLEAN	5	=	Transformed from int to Boolean.
Dimension	dim_animal	Freshwater	BOOLEAN	5	=	Transformed from int to Boolean.
Dimension	dim_animal	Aerial	BOOLEAN	5	=	Transformed from int to Boolean.
Dimension	dim_animal	Life.Habit.Method	VARCHAR(10)	5	=	Retrieved the data without any transformation.
Dimension	dim_animal	Mass.g	INTEGER	5	=	Transformed from BIGNUMBER to integer
Dimension	dim_animal	Mass.Method	VARCHAR(10)	5	=	Retrieved the data without any transformation.
Dimension	dim_animal	Island.Endemicity	VARCHAR(35)	5	=	Retrieved the data without any transformation.
Dimension	dim_animal	Diet.Plant	SMALLINT	5	=	Transformed from BIGNUMBER to SMALLINT.
Dimension	dim_animal	Diet.Vertebrate	SMALLINT	5	=	Transformed from BIGNUMBER to SMALLINT.

Continued on next page

Table 1: (Continued)

Dimension	dim_animal	Diet.Invertebrate	SMALLINT	5	=	Transformed from BIGNUMBER to SMALLINT.
Dimension	dim_animal	Diet.Method	VARCHAR(10)	5	=	Retrieved the data without any transformation.
Dimension	dim_animal	Genus.1.1	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	Species.1.1	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	Genus.1.0	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	Species.1.0	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	EltonTraits.1.0.Genus	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	EltonTraits.1.0.Species	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	IUCN.2016.3.Genus	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_animal	IUCN.2016.3.Species	VARCHAR(35)	4	=	Retrieved the data without any transformation.
Dimension	dim_towns_climate	Comunidad	VARCHAR(50)	2	=	Retrieved the data without any transformation.
Dimension	dim_towns_climate	Poblacion	VARCHAR(50)	2	Población	Retrieved the data without any transformation.
Dimension	dim_towns_climate	Latitude	VARCHAR(50)	2	Latitud	Retrieved the data without any transformation.
Dimension	dim_towns_climate	Longitude	NUMERIC	2	Longitud	Retrieved the data without any transformation.
Dimension	dim_towns_climate	Height_above_sea	NUMERIC	2	Altitud	Retrieved the data without any transformation.
Dimension	dim_towns_climate	Annual_Mean_Temperature	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Mean_Diurnal_Range	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Isothermality	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.

Continued on next page

Table 1: (Continued)

Dimension	dim_towns_climate	Temperature_Seasonality	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Max_Temperature_of_Warmest_Month	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Min_Temperature_of_Coldest_Month	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Temperature_Annual_Range	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Mean_Temperature_of_Wettest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Mean_Temperature_of_Driest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Mean_Temperature_of_Warmest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Mean_Temperature_of_Coldest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Annual_Precipitation	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Precipitation_of_Wettest_Month	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Precipitation_of_Driest_Month	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Precipitation_Seasonality	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Precipitation_of_Wettest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Precipitation_of_Driest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.

Continued on next page

Table 1: (Continued)

Dimension	dim_towns_climate	Precipitation_of_Warmest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_towns_climate	Precipitation_of_Coldest_Quarter	SMALLINT	3	=	Mapped the distinct climate feature for each town's coordinates.
Dimension	dim_dataset	Dataset name	TEXT	1	=	Retrieved the distinct datasets from which we obtained occurrences.
Dimension	dim_observer	Rights holder	TEXT	1	=	Retrieved the distinct observers that loaded an occurrence.
Dimension	dim_observer	Basis of record	TEXT	1	=	Retrieved the type of record associated with each observer.
Fact	fact_table_occ	Order	TEXT	1	=	Retrieved the data without any transformation.
Fact	fact_table_occ	Family	TEXT	1	=	Retrieved the data without any transformation.
Fact	fact_table_occ	Genus	TEXT	1	=	Retrieved the data without any transformation.
Fact	fact_table_occ	Species	TEXT	1	=	Retrieved the data without any transformation.
Fact	fact_table_occ	Date	TIMESTAMP	1	=	Change from String to Timestamp type. Dump the hour time and keep the date part.
Fact	fact_table_occ	Weekday	SMALLINT	1	=	Weekday calculated from Date Timestamp. Change from DOUBLE to SMALLINT.
Fact	fact_table_occ	Hour	SMALLINT	1	=	Hour time calculated from Date Timestamp. Change from DOUBLE to SMALLINT.
Fact	fact_table_occ	Longitude	DOUBLE	1	=	Retrieved the data without any transformation.
Fact	fact_table_occ	Latitude	DOUBLE	1	=	Retrieved the data without any transformation.

6 ETL Process

Insert figure here

The data sets involved in the project were collected from number of different sources explained above in the relevant section. All of these data sets have undergone an ETL(extraction, transform and loading) process. In this project, the ETL process has been automated mainly through Pentaho, but also a mix of R and Python.

The process finishes with the creation of the tables that constitutes the final star schema in our data warehouse in PostgreSQL. Namely, 'dim_animal', 'dim_towns_climate', 'dim_observer', 'dim_dataset', and the 'fact_table_occ' at the core of it. Both 'dim_animal' and 'dim_towns_climate' are meant to be filled once with the whole data on the animal species and towns in Spain, respectively. Meanwhile, the idea behind 'dim_observer', 'dim_dataset', and the 'fact_table_occ' is that they can be updated in a "regular basis" as the users add new occurrences into the database (or, in our case and given the extent of this project, to update it with new occurrences from *gbif.org*).This yields naturally to a setting in which we have 3 different transformations rather than a job that runs transformations sequentially: 2 independent ones to create and fill the first two tables (meant to be run once) and another one for the rest of tables (meant to allow for updates in the DW).

6.1 Stage 1: 'dim_animals'

As specified in the relevant section, Phylacine provides us with a csv format file that contains data about some relevant traits for all Mammals we have register of. Fortunately, Phylacine also provides us with a different csv file containing synonyms from other major biodiversity databases (as there is some debate on the correct nomenclature of certain species). This opens the door for a more robust occurrence ingestion system in which we can not only store one single type of nomenclature, but also compare with the rest of synonyms and, in case there is a coincidence with a synonym, allow for the ingestion of that occurrence in the proper format. Thus, our objective is to merge both the animal traits and the synonyms data for each of the registered species. Our Pentaho transformation would look like this:

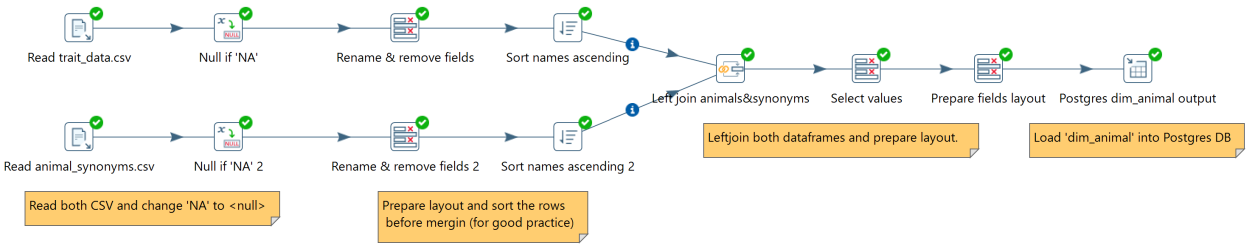

Figure 3: Pentaho transformation regarding the creation of 'dim_animal'.

6.2 Stage 2: 'dim_towns_climate'

We now aim to have a table that allows us to have an estimate on how the occurrences of each species are related to the climate information of that specific region. Obtaining climate data for each pair of coordinates of each occurrence is not only inefficient, but it also makes us rely on an external tool. Instead, we opt for an approach that will allow us to have the climate data with enough granularity to get a few insights as well as being independent from any internet connection: We will data for all the towns in Spain. Climate data doesn't change that abruptly, and the towns seem to cover all the surface with enough granularity. In order to make such table, we take advantage of WorldClim, and, via an API, we make a request by passing the coordinates of interest and this returns a .json file containing several variables of interest. The ETL process in this case is done

mainly via R as we can simply pass the coordinates retrieved from the csv file that contains the coordinates of each town in Spain, and we pass it through the API to obtain the Climate data we are interested on. Next, we merge both dataframes to obtain the definitive csv that we will finally pass to Pentaho to create 'dim_towns_climate'. A picture of the workflow in Pentaho can be seen in the next figure.

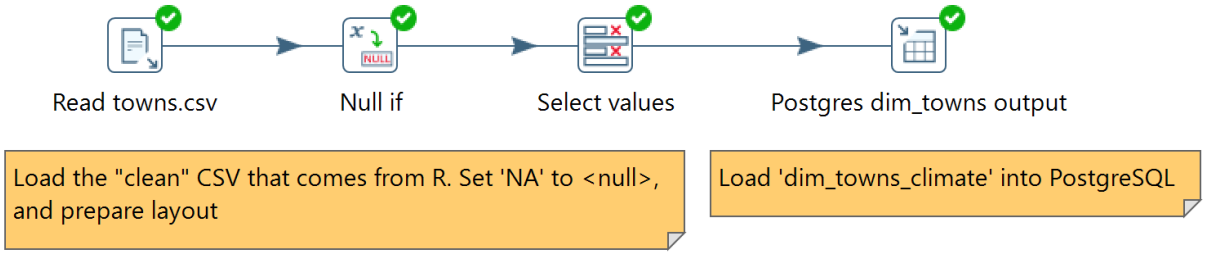

Figure 4: Pentaho transformation regarding the creation of 'dim_towns_climate'.

6.3 Stage 3: 'dim_dataset', 'dim_observer' and 'fact_table_occ'

Finally, we get to the final transformation of our ETL. This is the central hub in which all occurrences arrive (in this case, occurrences are taken from gbif database, but they might as well be user occurrences as long as they follow the mild conditions on the format. The transoformation can be devided into 3 major parts:

- On the first part we load the occurrences and we perform multiple comparisons with all the synonyms in our database to see if there's a match with any of the species registered. If so, the proper name of the species is passed along with the animal_id that will eventually establish a link with 'dim_animal'. The workflow can be seen in Fig. 5.

- On the second part, we use the calculator in order to retrieve the weekday and hour from the initial Timestamp of the occurrence. Next, we make use of the build-in Python interpreter in order to pass the coordinates to Cartesian format and calculate the distance to the closest town in Spain. Once obtained town_id and the distance in meters, we check whether it's close enough (14km) in order to infer the climatic variables (granularity condition). A picture of the workflow can be seen below in Fig. 6.

- Provided that the granularity condition is fulfilled, the name of the closest town is provided and the town_id will serve as a link to the climate table 'dim_towns_climate'. Else, the fields are filled with Nulls. Finally, the occurrences data regarding the observer and the source is divided into 'dim_dataset' (in case it's not already registered in there), 'dim_observer' (in case it's not already registered in there). Only when both dimension tables have been updated, it's time to register the occurrences into the fact table substituting the data about the observer/source by the respective id's that will link the fact table to the respective, this increases the efficiency of storage as we don't have to store the same names multiple times and allows the fact_table to contain only the necessary information. The workflow of this final part can be seen in Fig. 7

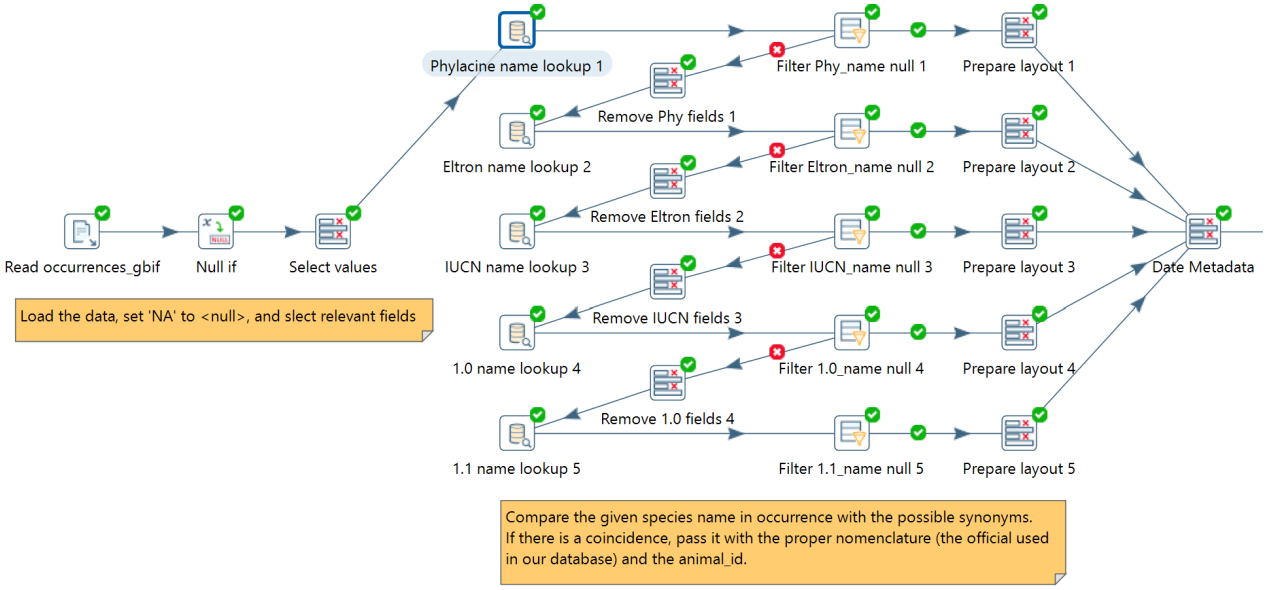

Figure 5: Part 1 of Pentaho transformation regarding the creation of 'dim_dataset', 'dim_observer', 'fact_table_occ.

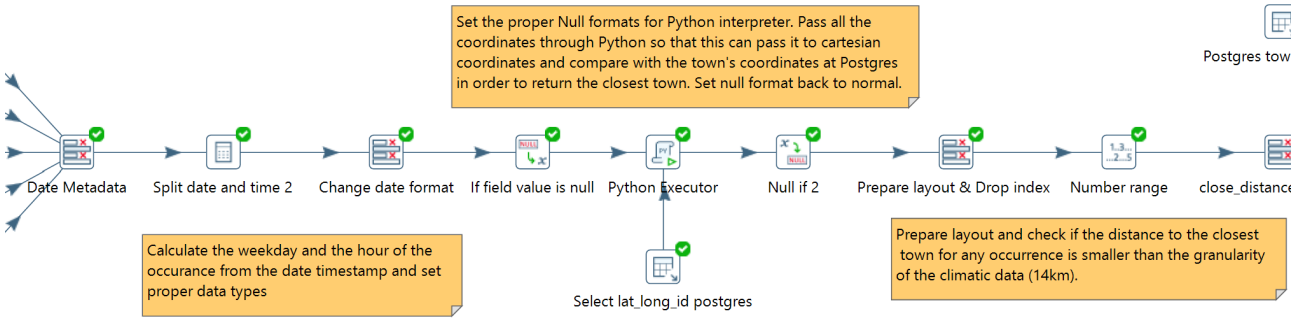

Figure 6: Part 2 of Pentaho transformation regarding the creation of 'dim_dataset', 'dim_observer', 'fact_table_occ.

7 Application

In this section we do some analysis on the data stored in our data warehouse, answering business intelligece questions related to the requirements listed in the introduction. These will be answered using QlikSense.

7.1 BI Query 1: How influential is the climate and elevation data in the number of occurrences?

The data sources we used for this question were *www.gbif.org*, *worldclim.org* and *www.businessintelligence.info/*.

The main findings are that there is no strong climatic influence regarding occurrences. The only

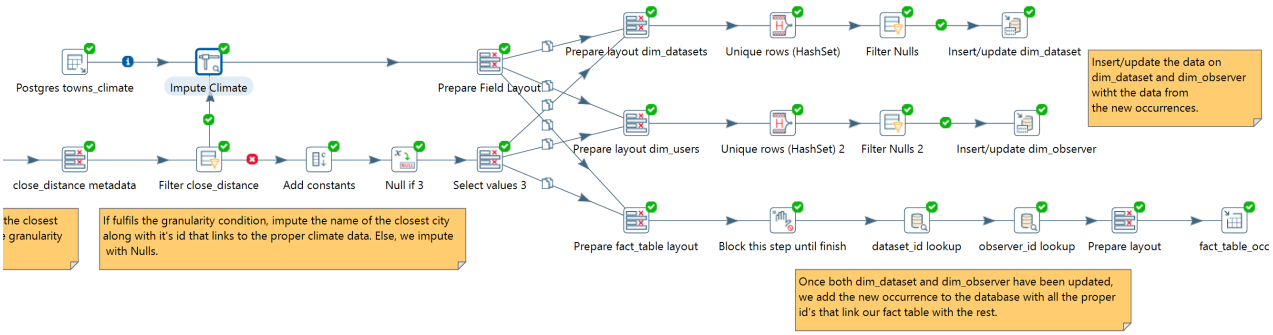

Figure 7: Part 2 of Pentaho transformation regarding the creation of 'dim_dataset', 'dim_observer', 'fact_table_occ.

relationship to be found is that the average elevation for towns with high occurrence is slightly mountainous. We identified that even though climatic variables had little visible effect on occurrences, there were certain regions with high occurrences and others with significantly less. This indicates that other factors affect occurrences such as regionality and the reach of the current occurrence infrastructure (Look at Fig 8, **??**, 10 and 11).

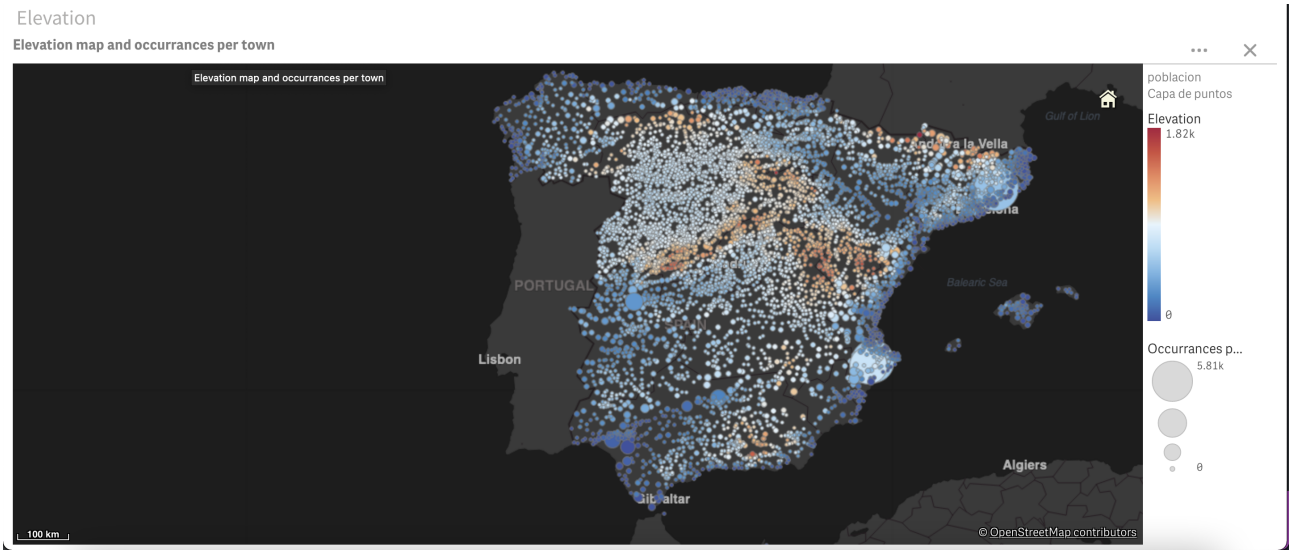

Figure 8: Elevation of town, size of bubble corresponds to number of occurrences

7.2 BI Query 2: How diverse is the pool of mammal carnivores in bigger regions within Spain?

The data sources we used for this question were *www.gbif.org*, *worldclim.org*, *https://megapast2future.github.io/PHYLACINE_1.2/* and *https://www.businessintelligence.info*.

We measured diversity not only by the number of unique species but by their diet percentages, mass and ecology. Any healthy ecosystem has diversity of all species, so the regions with the most species of *carnivora* are the healthiest. Unfortunately we see that in places with low occurrences we get high diversity and in places with high occurrences we get a somewhat lower diversity. We can see an impact in ecosystem complexity known high population centers (Look at Fig 12).

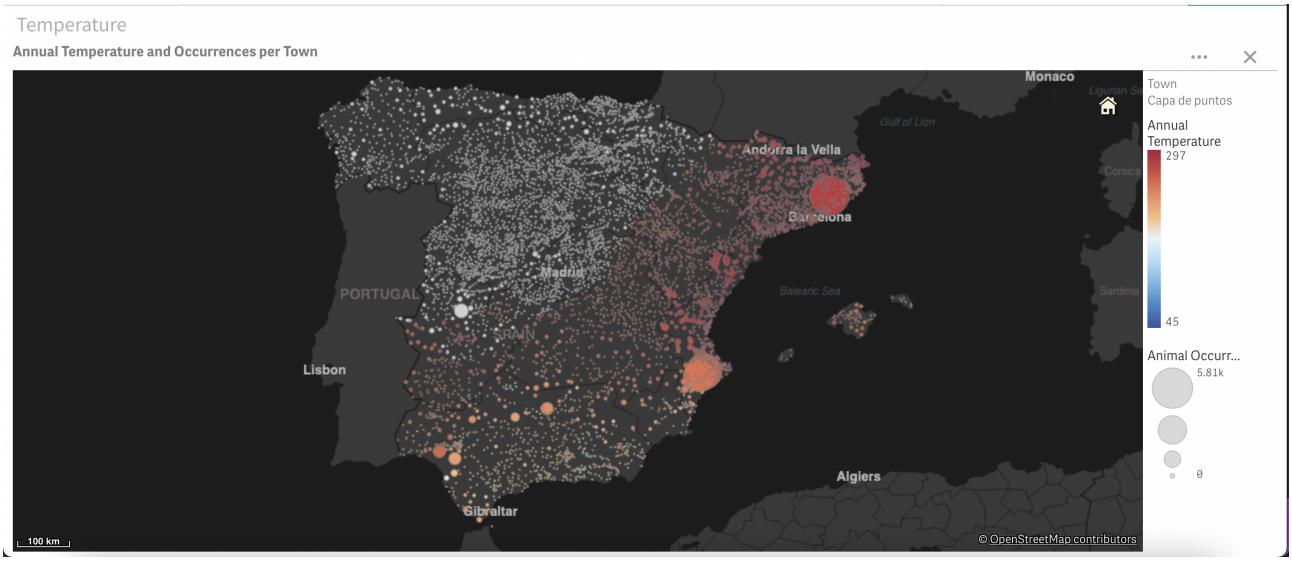

Figure 9: Average Annual Temperature by town, size of bubble corresponds to number of occurrences

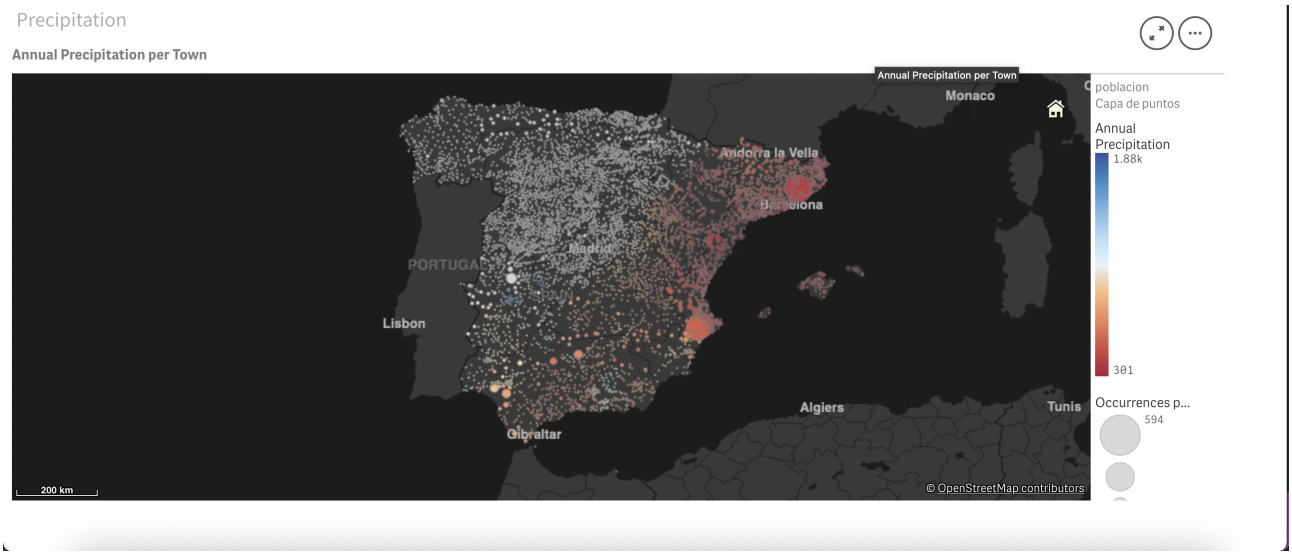

Figure 10: Annual Precipitation by town, size of bubble corresponds to number of occurrences

7.3 BI Query 3: Given the assumption that the highest quality of data is uploaded during business hours and on a work day, what is the quality of our data and who is uploading it?

The data sources we used for this question is *www.gbif.org*.

We can see that there is not a lot of information, most of our data has no information on the uploading time, it will be important to verify this information in the future. Of the non-random sample of data that we have we can see that: most uploads are on Tuesday and other weekdays, most of the uploads were made before 8 P.M. and after 10 A.M.

It is clear that the few data that we have hour and day information from is mostly high quality. We can also see that all of this data comes from just one Dataset: iNaturalist. We have seen before

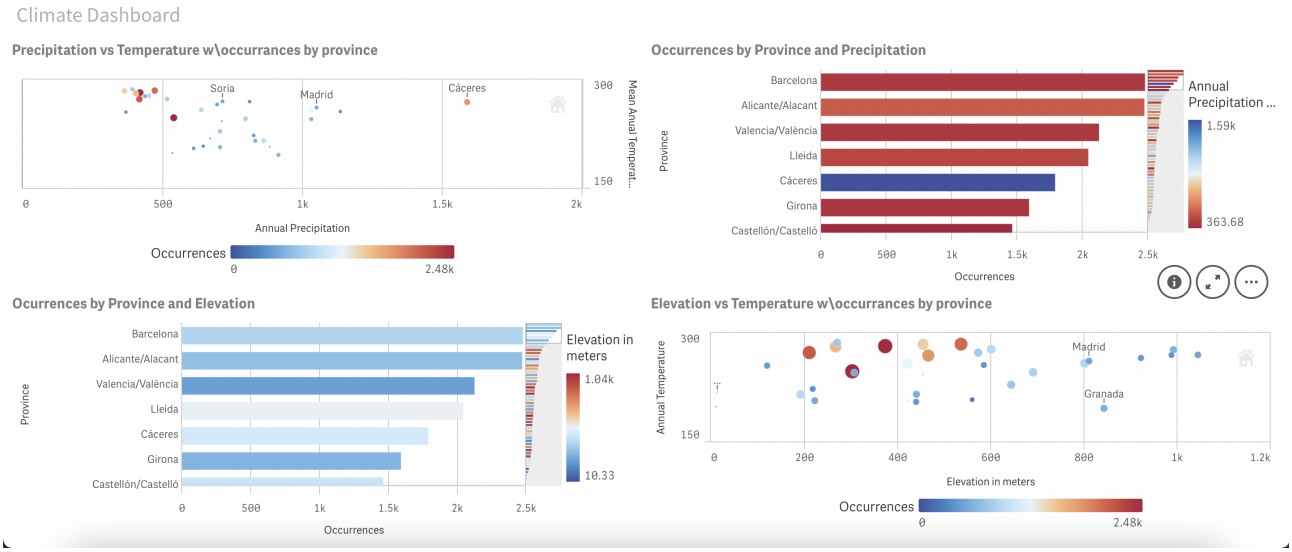

Figure 11: Climate Dashboard

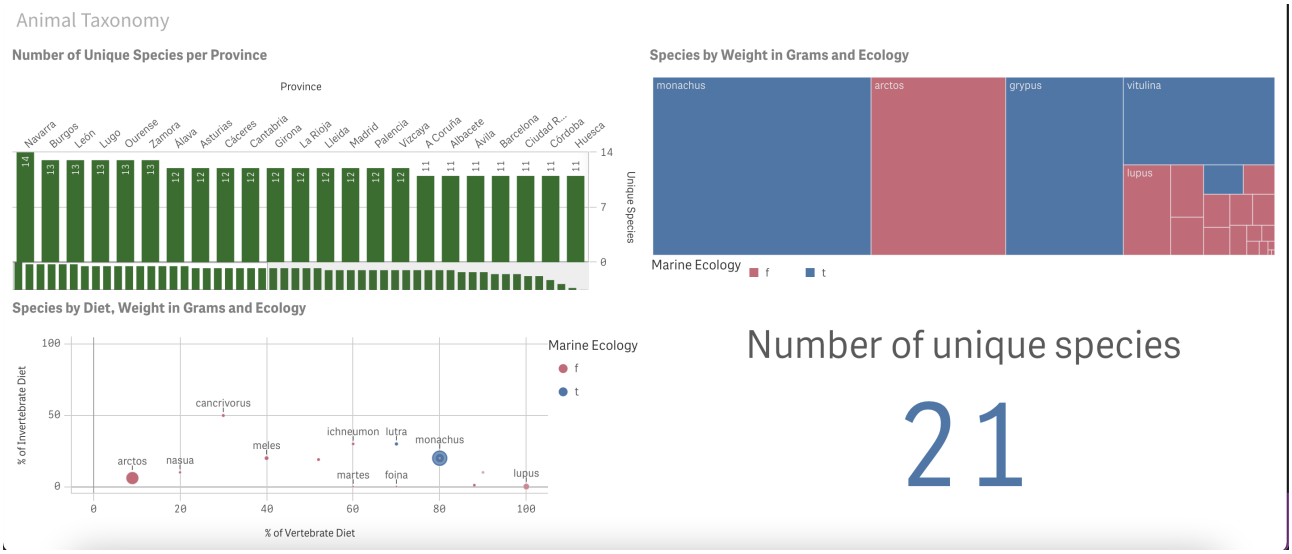

Figure 12: Animal Traits Dashboard, f = False and t = True

that this is a research grade Dataset, which corresponds to the high quality data assumption made earlier. At least from the non-random sample we are assured that some of our data is high quality, we can monitor future uploads and see if it corresponds to our high quality data assumption (Look at Fig 13).

7.4 Discussion

The need to understand biodiversity, ecology and the influence of human activities in our environment is crucial for future research. Using the diversity of the order *carnivora* as a proxy for ecological health, we can extract useful information from occurrence data sets.

We understand that even though ecological conditions are linked to climate, occurrences are

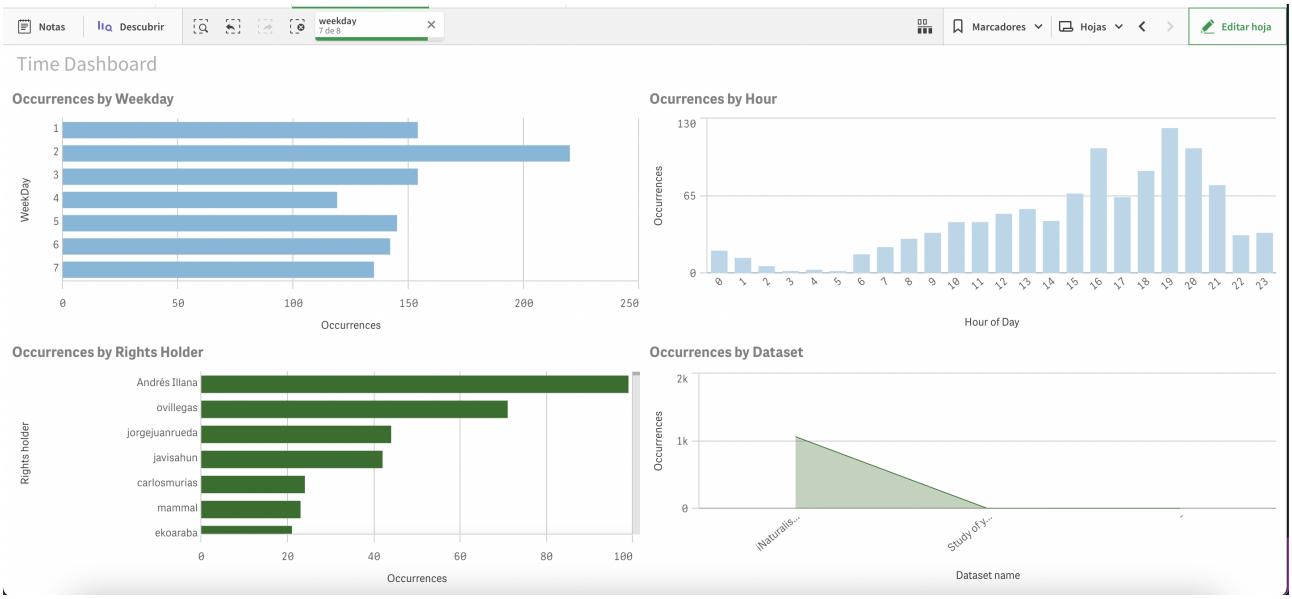

Figure 13: Observer and Time Dashboard

more linked to human population centers with active research uploaded to *www.gbif.org*. This indicates a new direction to have a complete view of occurrences in Spain for the future. We also identified that there are not too many unique *carnivora* in Spain, but they are very diverse in diet and size, which is a great indicator of diversity of prey.

We also got positive feed on our hypothesis (assumption) that time data is crucial for understanding the quality of our data. Linking all of these results together we have a good view of the data at hand but we also have a promising road map for the future, given the missing and incomplete data we observed.

8 Future Work

This project represent an MVP of what we wish to do. We want to create an platform from which people, institutions and private research companies can upload their information and keep track of occurrences. We also allowed for growth, meaning that we have left many unused variables available in our infrastructure to allow for new countries, species, islands, climate and information about animal traits. Having these sights clear we can provide the next steps in our journey.

To see the whole picture we need to fill in the missing data, specifically climate data for the north-western half of Spain. For this we could scrape data from local weather stations, but this does not solve the lack of observations in what is considered by experts as one of the most bio-diverse regions in Spain. For this second issue we must focus on reaching a wider audience that stretches from east to west and north to south. More observations will give us a better picture of the bio-diversity of Spain. Lastly, if we want to get an estimate of the quality of our data we need time stamps on every upload. This would give us a good proxy for occurrence quality for future use.

References

[1] W. J. Bond. Keystone species. In *Biodiversity and Ecosystem Function*, pages 237–253. Springer Berlin Heidelberg, 1994.

[2] Reed F. Noss, Howard B. Quigley, Maurice G. Hornocker, Troy Merrill, and Paul C. Paquet. Conservation biology and carnivore conservation in the rocky mountains. *Conservation Biology*, 10(4):949–963, August 1996.