Monkey Business:

Developing a Taxonomy of Primate Observations

A research project for Knowledge Representation on the Web

Alex Hoorn (2716056)
Bram Kreuger (2712191)
Nathan Vaartjes (2596846)
Sunny Soni (2705047)
Yuyu Bai (2732696)

Group 6 Vrije Universiteit Amsterdam

September 10, 2023

Abstract

In this paper, we present a comprehensive ontology of the world's primates, excluding the Homo genus. The motivation for this research was to test the feasibility of combining and integrating various open source resources and tools for knowledge graphs. This ontology is based on the RDF-ized NCBI taxonomy, and is populated with iNaturalist observations. It includes a variety of open source data sources, including EOL, GBIF and others. The ontology employs the Darwin Core standard vocabulary, which provides a standard and facilitates reuse for researchers. We created a web application to create insight into the resulting ontology. The ontology and web application are freely available to anyone interested.

1 Introduction

Modern technology such as knowledge graphs (KG's), paired with data visualization can give us great insights into information which was difficult to see an overview of before. KG's are the perfect tool for an application where you want to combine various resources into a single collection. The hierarchical structure is ideal for animal taxonomies where families have genera which again have multiple species, and so forth. In a traditional structured databases retrieving this information would require multiple operations, but in a KG it is possible to infer this information with a single simple query. Inferencing provides value to a KG approach, by leveraging logic rules to derive implicit knowledge in a graph. As such, instances belonging to a distant superclass can again be queried simply via that superclass, using the transitivity of the subclass relation.

To this end we decided to create an ontology consisting of the world's primates (minus the Homo genus, which includes humans). The primate population has been under severe pressure in the last years [10]. To gain insights in how the population is developing and where the regions are where they are under pressure the most, we have created this ontology. Besides giving some interesting insights in the spread of primate species, this paper can be viewed as a demonstration of how various open source tools, ontologies, datasets and vocabularies can be combined into a working example of an ontology which is accessible with a custom web app. Once the ontology is created, users will be able to run queries like: "Where can the orangutan be found", "What is the biggest spread of the Chimpanzee species" and "Which periods yielded the most observations for specific taxa". Some of the challenges we ran into in the paper is the integration of the various resources we used. Since not every resource uses the same format it is a challenge to integrate these into our solution. In this paper we try to answer the following question: "Is it possible to use various open source tools, vocabularies and openly available datasets to create a taxonomy which can be accessed by a web application?" Since this is mainly a development project, the paper will have a focus on the approach and methodology.

The structure of the paper is as follows: We start of with the related work in section 2 where we will go through research papers, describing the various tools and resources we used. We then move on to the methodology 3 where we explain how we obtained our data and how we used it to create our ontology. The observations subsection 3.1 goes into depth how the primate observations were retrieved from iNaturalist and how they were converted into the right format. The locations subsection 3.2 discusses how the geospatial data in the iNaturalist data is linked in our dataset using GeoNames. The next subsection, ontology design 3.3, discusses how NCBI was used to obtain a taxonomic classification ontology. In the following subsection, integration 3.4, we discus how the NCBI graph got integrated with our data. In publishing 3.5, we explain how we published our ontology using Triply and Github. After this we move on to the results 4 where we discuss how we created an web app that makes it possible to

interactively explore aspects of the ontology. Finally, we have the conclusion 5 where we give a brief overview of what we did in this project.

2 Related Work

In this section we discuss various papers which are related to our project or which introduce tools or resources which are used in our project.

GBIF

In the paper "Spatial bias in the GBIF database and its effect on modeling species' geographic distributions" [4] the authors model a species distribution. This is similar to the task we have at hand. In this paper the authors investigate the spatial bias which originates because of an uneven effort in sampling, data storage and mobilization. This specific research topic, interesting as it is, is not directly applicable for our research. However, the task which precedes the investigation of the bias is interesting to us; modeling the geographic distribution of a species. In the paper they use a common Eurasian butterfly as a taxon for this model, while we use primates; this distinction does not matter for the process. We got some interesting insights into their methods by reading this paper.

GeoSPARQL

The paper "GeoSPARQL: Enabling a Geospatial Semantic Web" [3] introduces us to the new semantic web standard: GeoSPARQL. With GeoSPARQL it becomes possible to use geospatial data contained in linked open data using SPARQL. With this tool, users can query things like: "Which municipalities border the city of Amsterdam". This is relevant for us because we want to be able to query similar things with regards to our primate taxonomy. Once we obtained results from this query we want to be able to visualize these results. To do this we need the geospatial results to feed to our web application.

iNaturalist

In "iNaturalist as an engaging tool for identifying organisms in outdoor activities" [12] the authors evaluated a new and upcoming smartphone application. The iNaturalist project enables individuals all over the world to upload their pictures of animals to the website, along with the location where it was observed and other metadata. The animal taxon ID is then identified through a community-driven process, where individuals can suggest ID's and others can review them. Given the popularity of the platform, using this data allowed us to use a large quantity of up-to-date data of good quality.

NCBI

The paper "The NCBI Taxonomy database" [6] the authors propose a taxonomy database which includes organism names and taxonomic lineages. It contains many intertwined resources from the NCBI. Some of the data in the NCBI database comes from the Catalogue of Life, the Encyclopedia of Life, NameBank and WikiSpecies. The resulting database is very useful to our project since it contains a clearly defined species taxonomy. The NCBI taxonomy also holds a referential status in terms of quality. To this end we used the database for its subclass relations for the primate species which could then be used to order the data we obtained from other sources.

GeoNames

GeoNames[1] is a well-known geographical dataset. It originally is a collection of tabular datasets that describe geographical locations. But after the introduction of the paper "A semantic schema for GeoNames" [9] it has also been converted to a knowledge graph. It currently contains over 12 million unique features [1] integrating information such as latitudes/longitudes, population and location types. Additionally it has relationships for every location so that, for example, the city of Amsterdam is part of the country The Netherlands. Using this information we can relate the observations to regions, countries, continents and more.

DarwinCore

Having a consensus on terms used to describe an observational dataset is crucial with regards to data reuse and knowledge sharing. Darwin Core is a standard vocabulary of terms [13], used to describe biodiversity datasets. For example, `dwc:observedOn` is used to universally denote on what date a certain animal has been observed within a certain dataset. Especially in the context of Linked Open Data, where data reuse and linking is the core idea, having a standard vocabulary is of great value. We decided to use DarwinCore terms when converting our observations to RDF.

Building upon these works, we developed an application allowing users to interactively visualise primate location data on a map through SPARQL queries, or browse through preconstructed queries. This allows us to leverage the power and flexibility of the RDF and OWL languages to get new and precise insight into real-world data.

[1] https://www.geonames.org/

3 Methodology

This section describes the process of extracting and integrating primate observations, as well as the process of ontology design, creation and publishing.

3.1 Observations

To retrieve primate observations, we decided to use iNaturalist observations, available from https://www.inaturalist.org, accessed [15-03-2022]. In order to have the observation records locally, we extracted iNaturalist observations via their API. At the same time, we also extracted the same research grade data on primates from iNaturalist using GBIF's export tool [7]. Every observation record contained metadata about its observation location. Using a SPARQL query dump from WikiData containing NCBI ids of all iNaturalist Observations, we combined them with our original iNaturalist extracted data. This yielded a large CSV file. This file was then converted to RDF via the OntoRefine tool of GraphDB. To link observations to their attributes such as the observation date, we used DarwinCore terms. It must be noted that because the iNaturalist data format as well as the OntoRefine mapping from CSV to RDF are static, it is trivial to update the observation database automatically, were this tool to be put to use. Figure 1 shows an example of how observations are represented in the graph. The JSON mapping used to convert from CSV to RDF as well as the original CSV file and any other file related to the process can be found on our Github[2].

3.2 Locations

The observations data from iNaturalist contains geographical latitudes and longitudes. However, it does not contain any named information of the location such as the country or its type of location like a mountain or forest. To be able to do this we linked our observations with GeoNames.

The full dataset of GeoNames is available as an ontology through its service, which can output a simple get query of a single object to RDF. But it lacks a proper SPARQL endpoint, making navigating and linking the ontology a difficult task. The ontology is also available as a direct file download. So its possible to locally host it with a service providing a SPARQL endpoint. However, the downloadable ontology is over 17 gigabytes, making it infeasible to run on our available hardware.

To tackle these problems we decided to partially reconstruct the GeoNames ontology containing only the information we are interested in, using the available tabular datasets. We only kept types, names, geographical locations for objects of types such as countries, regions, lakes, parks, cities, mountains and forests. We linked these up with their hierarchical relations. This way we recreated the GeoNames ontology specifically for our use in only 100 megabytes. Additionally, we were able to add `geo:partOf` as a `owl:TransitiveProperty`. So now we can infer that if `<Amsterdam partOf North_Holland>` and `<North_Holland partOf The_Netherlands>` then also `<Amsterdam partOf The_Netherlands>`.

Using our smaller GeoNames ontology we were able to link up the iNaturalist observations to a named location. We did this by calculating the nearest geographical point for every entry using the haversine function. This formula (1) is able to calculate the distance between two geographical points of latitudes and longitudes. It is able accurately calculate distances over our spherical earth [11].

$$D(x, y) = 2 \arcsin \left[\sqrt{\sin^2((x1 - y1)/2) + \cos(x1) \cos(y1) \sin^2((x2 - y2)/2)} \right] \tag{1}$$

3.3 Ontology Design

In parallel, we needed to find a taxonomic classification ontology, to organize the observations into a framework. Given that exact taxon delimitations and their hierarchy are still debated, we opted to base our ontology on an RDF-ized version of the NCBI taxonomy [2], given the referential status of the NCBI. Furthermore, the iNaturalist taxonomy backbone relies on multiple established taxonomies, of which the NCBI taxonomy is one of the biggest. A large part of our observations (approx 60%), contained the NCBI taxon ID, which greatly simplified the task of matching the observations to the ontology.

The ontology is organized as every taxon being a class. These classes are linked via subclass relations in accordance with their taxon hierarchy. Every class has a link to its taxon rank, such as "subfamily". Importantly, the ontology contained no instances. We therefore decided to define our observations as instances of the appropriate classes.

3.4 Integration

The NCBI graph as well as the observations graph were loaded into GraphDB, and were merged via a SPARQL construct query. This query can be consulted on the project's Github repository. Every observation was linked to the corresponding taxon class with `rdf:type`, using the NCBI ID. Observations with no ID were discarded (approx. 40%).

[2]https://github.com/AlexHoorn/PrimateObservationsOntology

Additionally, all taxon classes were aligned to equivalent taxa from other naming authorities via a SPARQL service query on Wikidata. Wikidata keeps a reference of equivalent class ID's across naming authorities (for example, NCBI ID 9606 = ITIS ID 180092). Naming authorities considered here are Encyclopedia of Life (EOL), iNaturalist, ITIS and GBIF. The classes were linked with `skos:exactMatch`. We decided to not use `owl:equivalentClass` nor make the observations members of classes from other naming authorities, as these are not ontologies, but rather structured datasets. As such, these URL's are not technically classes, and their URL's could change. We therefore decided not to base inferencing on them at the moment. However, we did think adding `skos:exactMatch` would add information that can still be used in queries, without creating potential problems.

In total, we have mapped our ontology to 9 other ontologies or structured datasets (INAT, WIKIDATA, DCTERMS, ITIS, EOL, GBIF, DWC, NCBI, GEONAMES), and it contains 1,168,173 skos:exactMatch links. A graphical summary of our ontology and observations can be found in Figure 1.

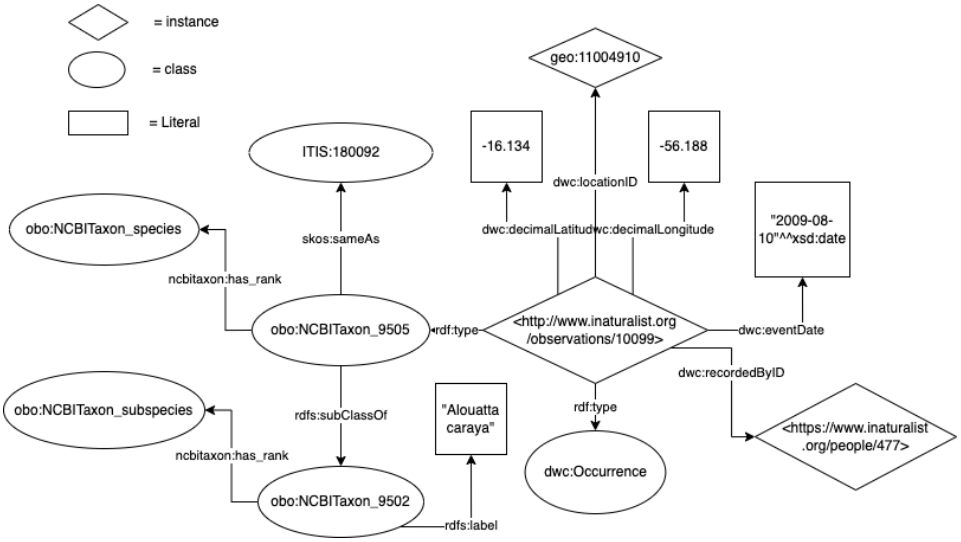

Figure 1: Ontology diagram

An additional advantage that the NCBI taxonomy ontology provided, is the inferred information it contains. The transitivity of the rdfs:subClass relation for example, enabled us to fetch all observations of subclasses of a given class, such as all the lemurs, while the class lemurs itself is only a taxon and contains no direct observations. In total our ontology contains 1,708,593 explicit triples. With RDFS reasoning the inferred triples amount to 2,742,533, resulting in an expansion ratio of 2.61. With full OWL reasoning the ontology grows by 5,954,874, resulting in an expansion ratio of 4.49. This reflects the advantage of using RDF KG's for insight creation in general.

Documentation triples such as `dc:tile`, `dc:creator`, were added manually to the ontology. A documentation was then automatically extracted via pyLODE [5]. Both a short and a full version (containing all 15052 class descriptions) can be consulted on our Triply page[3].

3.5 Publishing

We have used Triply provided by Vrije Universiteit Amsterdam[4] as our triplestore. This makes our ontology publicly available with a SPARQL endpoint that has Jena RDFS reasoning enabled, so as to be able to retrieve all the RDFS inferred triples in the graph. However, in our testing some of our queries frequently timed out. We assume this occurs because the ontology is quite large and has many inferred triples. So the server might not be able to process every query in a timely manner. As a backup, we have published our ontology on Github[5], so that anyone can open the ontology with a locally hostable triplestore such as GraphDB.

To be able to easily provide examples and results of queries to our ontology we built a web application. This application uses Python with Streamlit[6]. Streamlit is a package providing the tools to quickly build a data-driven application including visualisations. With this we have created some tools to interactively query and visualise possible research aspects of our ontology.

4 Result

An interactive web application was created with the help of Python and Streamlit to make the data analysis and exploration easier. This web application makes it easy for wildlife researchers and institutions to browse and analyze

[3]https://krr.triply.cc/NathanV/KRWprimatestaxonomy/assets
[4]https://krr.triply.cc/
[5]https://github.com/AlexHoorn/PrimateObservationsOntology
[6]https://www.streamlit.io/

our ontology. This program may be used to analyze additional observations of primates if they become available in the future, by updating the ontology powering the application. The app's primary features are described in this section.

4.1 Data Summary

The app's home page is a data overview page, as depicted in Figure 2. We see a basic overview of the ontology on this page, including the total number of observations, the total number of taxa, and the total number of locations. All the locations of observations are represented graphically on the map. This map is accessible in two styles in our app, the first being Cartographic and the second being a Satellite view, allowing the user to see the information with both clearly defined regional borders or natural features. We may acquire a visual representation of the current data using this display.

It is immediately apparent that primate species can be found around the equator and in the southern hemisphere, in regions such as South America, Africa, and southern Asia. This distribution matches what we already know. We can deduce from this graph that the species' distribution is linked to climate, as the primate species are found in hotter areas. This image depicts three "hotspots", all of which, as the graphic shows, are essentially around the tropics. The density of primates is high in these three areas. The distribution density may also be related to the regional economic situation, as we note that this has a similar pattern to the Human Development Index [8]. This is currently only a conjecture, and should be investigated by comparing it to other variables.

Figure 2: Observations summary page

4.2 Map of Taxon Observations

The app's second page can visualize the precise locations of the observations. Giving a hint about the distribution of specific taxa by different ranks. Here you can select a certain rank from the hierarchy. The number of observations for each corresponding taxon will be displayed in the histogram. In our example shown in Figure 3, there are seven families, and the map below depicts the geographical distribution of observations for each family. It is possible that the species with the most current observations also has the largest population. But note that the number of observations also relies on other external factors such as the accessibility of the living area of a taxon.

The example page, at Figure 3, shows the number of observations and geographic distribution of various ranks, such as certain families. Cercopithecidae is primarily distributed in Asia and Africa, as shown in the geographical distribution map below. However, interestingly some are found in the Caribbean. The Atelidae family is only active in South and Central America.

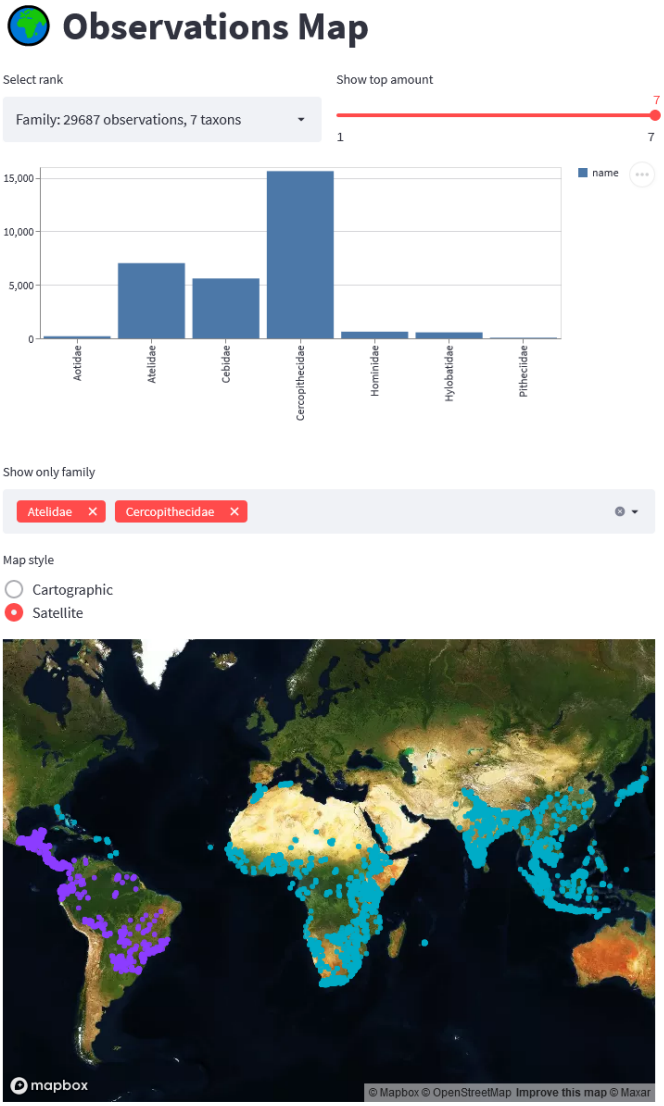

Figure 3: Observations map page

4.3 Spread Comparison

This section of our app interface has been built to compare the distribution of two taxa on the basis of the distance between their most distant observations. Due to heavy computational and space requirements of the calculations, we have only selected the ranks with an amount of corresponding taxa greater than ten. For example, we can compare the distributions of two genera that we are interested in. The farthest distance between two observations from the Alouatta genus and Brachyteles genus from all observed locations are 7400.28 Km and 411.215 Km, respectively, as seen in Figure 4. If sample bias is not taken into consideration, this suggests that the Brachyteles genus has a much smaller living area than the Alouatta genus.

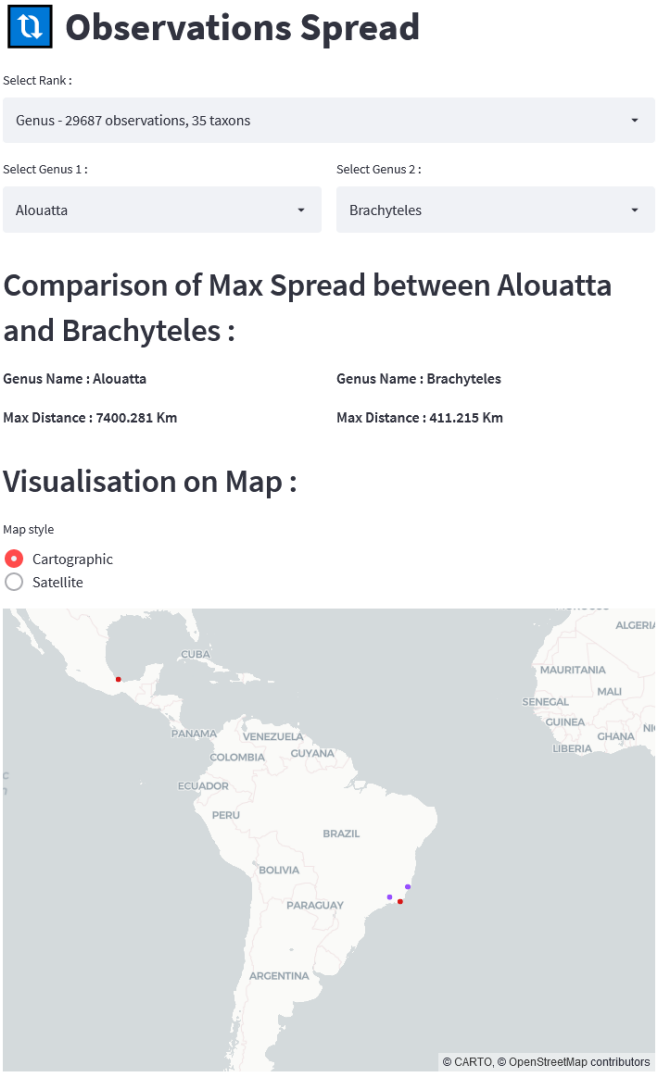

Figure 4: Observations spread page

4.4 Trend of Observations

One of the values that we are most interested in while studying wildlife is the change in the number of wild animals observations made. We also considered the function of each taxon's trend over time in our application study.

One of the intriguing things we saw was that once the COVID pandemic began in 2020, there was a steep decline in the amount of observations of primates. The new observations for each taxon decreased in 2020, before increasing in 2021. We assume that this is related to researchers' restricted activities and the temporary decrease in tourism. The program also allows us to track how each taxon's observation counts have evolved over time. This could be very useful for studying these taxa.

For example, the genus Colobus, for which there are just limited observations in the single digits, could be a scarce or endangered species. During the pandemic, the overall recorded values declined, limiting human activity and maybe allowing for an increase in the population of specific species. The species Alouatta guariba has seen a sudden and considerable increase in numbers since 2019, possibly indicating that human activities has had a big impact on this species. In contrast, from 2019 onwards, the Aotidae family has seen a decrease in observations, which could possibly (according to our limited general domain knowledge) indicate one of the few things such as: the species' survival is dependent on human protection, there are less observations as a result of limited observer activities during the pandemic, etc.

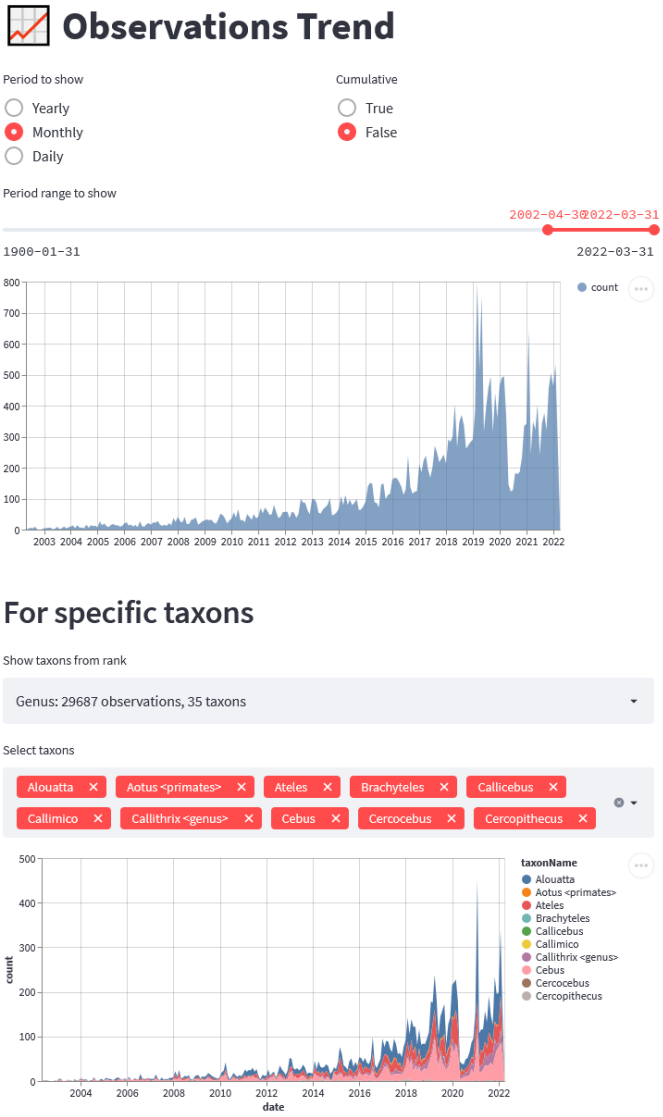

Figure 5: Observations trend page

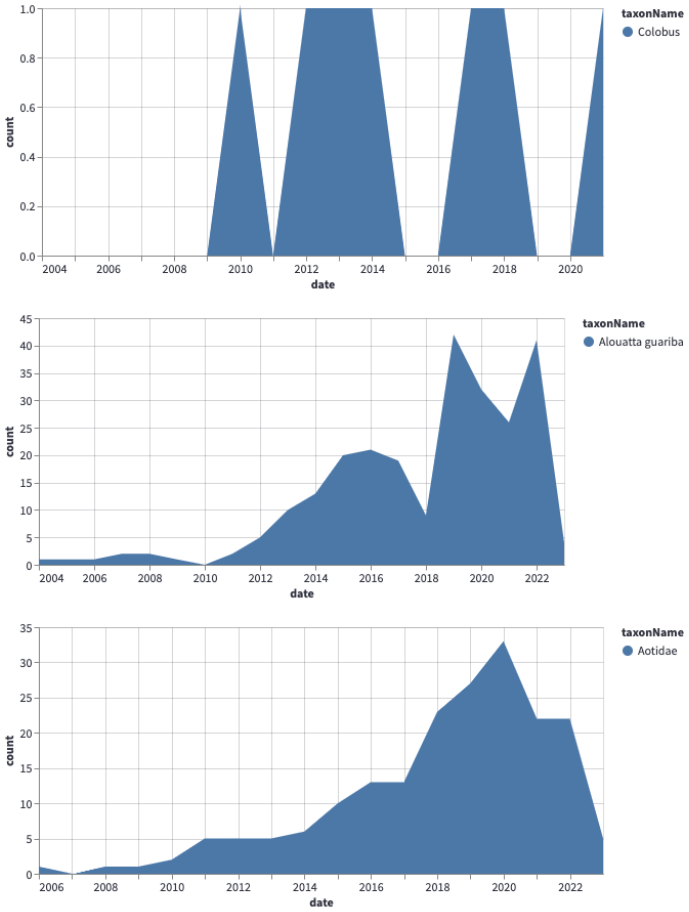

Figure 6: Observations for some specific taxa

4.5 SPARQL Endpoint

We have also included a SPARQL query interface (Fig 7) in our web app to give the researchers the possibility of using their own queries to search our knowledge graph for the information they require. This enables the users of our app to query our ontology based on their own interests. Our web app's procedural implementation in other pages is also largely dependent on some of SPARQL queries in the backend. The SPARQL capabilities of our ontology are described in details in the following section 4.6.

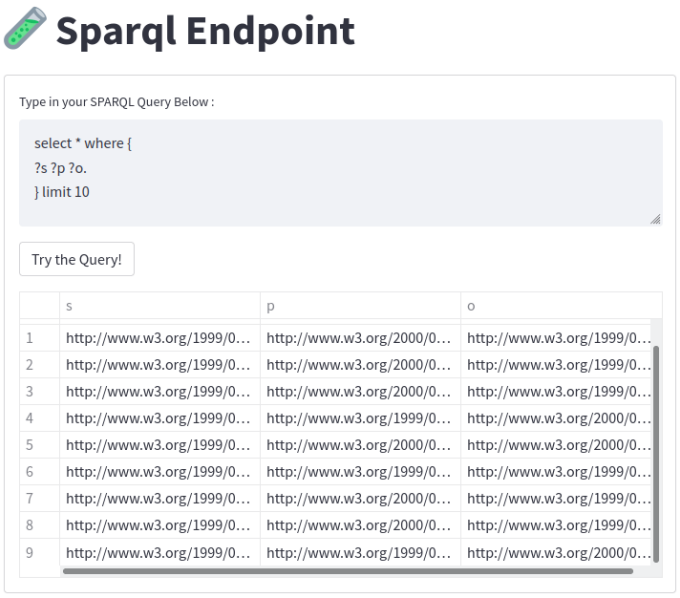

Figure 7: SPARQL Endpoint Interface

4.6 SPARQL Query Capability

Given the way our ontology and knowledge graph has been designed, using SPARQL queries, the users can extract information from it. The users can not only query using our data, but they can also query using our data in conjunction with other databases to obtain the information they seek. Using an example of the query we have used for building one part of our web app, you can see the demonstration of this capability in action. Using the query given in Listing 1, we have been able to extract the observation data from our graph based on their division by genus, which were inferred using the subclass relations from the NCBI ontology. By connecting it to DarwinCore based properties in the query itself, the extraction of the positional data of observations on the basis of their genus was made possible. Additionally, using the subclass relations we are able to quickly calculate the amount of observations for every rank, even if its corresponding observations do not have an explicit direct link to the rank, as shown in Listing 2.

Listing 1: SPARQL Query for fetching observations data based on ranks

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ncbitaxon: <http://purl.obolibrary.org/obo/ncbitaxon#>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
SELECT * WHERE {
    ?obs a dwc:Occurrence ;
        a ?kind .
    ?kind ncbitaxon:has_rank <http://purl.obolibrary.org/obo/NCBITaxon_genus> ;
        rdfs:label ?name .

    OPTIONAL{?obs dwc:decimalLatitude ?lat}
    OPTIONAL{?obs dwc:decimalLongitude ?lon}
}
```

Listing 2: SPARQL Query for counting and filtering the amount of observations per rank

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ncbitaxon: <http://purl.obolibrary.org/obo/ncbitaxon#>
PREFIX dwc: <http://rs.tdwg.org/dwc/terms/>
SELECT ?rank ?rankLabel
    (COUNT(?rank) as ?count)
    (COUNT(DISTINCT ?kind) as ?kindCount)
WHERE {
    ?obs a dwc:Occurrence ;
        a ?kind.
    ?kind ncbitaxon:has_rank ?rank .
    ?rank rdfs:label ?rankLabel .
}
GROUP BY ?rank ?rankLabel
HAVING (?kindCount > 1 || ?rankLabel = "order")
ORDER BY DESC(?count) ASC(?kindCount)
```

4.6.1 Qualitative Analysis

Due to the consistent nature of our ontology, different kinds of operations like the one explained above are possible using the SPARQL backend, depending on the needs of the query. In terms of qualitative analysis, the capabilities demonstrated by our web app as well as the creative possibility of extracting data using the SPARQL interface provides some positive feedback to the utility of our research.

It must be noted, however, that the capabilities of reasoning-enabled SPARQL come at a price, which becomes apparent using the SPARQL endpoint on Triply. When working with larger datasets, this issue might be amplified and ultimately defeat the purpose of KGs. It is therefore crucial to have efficient queries and a solid triplestore for ontology-based approaches such as a locally hosted GraphDB server to be useful in practice.

5 Conclusion

Diverse data sources on the internet often contain different information and are available in various data formats. When conducting research on a topic, it is common for developers to have to integrate data from many sources.

We used data from a variety of open sources for this project, including iNaturalist, GBIF, NCBI, and others. We were able to incorporate these data sets to construct a comprehensive knowledge graph for observations of primates. We have published this ontology on Triply, so any researcher with an interest in these observations can openly access it. The ontology employs the DarwinCore standard vocabulary, which provides a standard and facilitates reuse for researchers.

We have also built a web application that interactively shows the information in our ontology in several detailed visualisations. Users can not only retrieve the data they're interested in, but also track the geographic distribution and observation fluctuations of individual taxa over time with this tool. We also interpret some noteworthy phenomena in the visualisation interface in this paper.

Overall, the project is quite valuable. Not only can people interested in learning about KG's learn about the entire process of collecting data, integrating data, building ontologies, publishing data, and visualizing using a web app, but wildlife researchers can also get a good example of using KG's to investigate animal distribution, migration, and racial changes.

A more complete and comprehensive KG is a possible future work. iNaturalist collects observations of the whole Animalia kingdom. But in our paper, we have focused primarily on the Primates order. Additionally, we opted to base the taxon hierarchy on NCBI. However, the true biological hierarchy is still debated. The database Catalogue of life, for example, also contains a complete structure of taxa that could differ and be taken into account in the future. Finally, whilst we have relational GeoNames information available in the ontology, we have not actively used this in our research. This may still provide useful insights for some future research. The distance calculations done in Python can be offloaded onto the SPARQL server with the proper use of GeoSPARQL queries using our knowledge graph.

References

[1] About geonames. https://www.geonames.org/about.html.

[2] Frederic Bastian. Ncbi organismal classification. https://obofoundry.org/ontology/ncbitaxon.html.

[3] Robert Battle and Dave Kolas. GeoSPARQL: Enabling a geospatial semantic web. page 17.

[4] Jan Beck, Marianne Böller, Andreas Erhardt, and Wolfgang Schwanghart. Spatial bias in the gbif database and its effect on modeling species' geographic distributions. *Ecological Informatics*, 19:10–15, 2014.

[5] Nicholas Car. pylode: An owl ontology documentation tool using python, based on lode. https://github.com/RDFLib/pyLODE.

[6] Scott Federhen. The NCBI Taxonomy database. *Nucleic Acids Research*, 40(D1):D136–D143, 12 2011.

[7] Occdownload Gbif.Org. Occurrence download, 2022.

[8] Jack Harris, Collin Ritchie, Glenn Hanna, Joseph McCain, and Yisi Ji. The inequitable global burden of lip and oral cancers: Widening disparities across countries. *Journal of Oral and Maxillofacial Surgery*, 79, 12 2020.

[9] Vincenzo Maltese and Feroz Farazi. A semantic schema for geonames. 2013.

[10] Russell A Mittermeier, Janette Wallis, Anthony B Rylands, Jörg U Ganzhorn, John F Oates, Elizabeth A Williamson, Erwin Palacios, Eckhard W Heymann, M Cecília M Kierulff, Long Yongcheng, et al. Primates in peril: the world's 25 most endangered primates 2008–2010. *Primate Conservation*, 24(1):1–57, 2009.

[11] C. C. Robusto. The cosine-haversine formula. *The American Mathematical Monthly*, 64(1):38–40, 1957.

[12] Shem Unger, Mark Rollins, Allison Tietz, and Hailey Dumais. inaturalist as an engaging tool for identifying organisms in outdoor activities. *Journal of Biological Education*, 55(5):537–547, 2021.

[13] John Wieczorek, David Bloom, Robert Guralnick, Stan Blum, Markus Döring, Renato Giovanni, Tim Robertson, and David Vieglais. Darwin core: An evolving community-developed biodiversity data standard. *PLOS ONE*, 7(1):1–8, 01 2012.

A Annex

A.1 Contributions

In this section we briefly explain which parts of the report and the project we each contributed to.

- **Alex Hoorn** Did the recreation of part of the GeoNames ontology and mass matched the observations to it based on distance. Build a portion of the web app (home, map, trends) and all of its underlying SPARQL queries. Wrote the GeoNames/locations portions and small pieces all around.
- **Bram Kreuger** Initially explored iNaturalist. Worked on exploring different methods to extract the ontology from NCBI and COL. Did most of the literature research and wrote the according section. Finally wrote the introduction and the abstract.
- **Nathan Vaartjes** Contact with Lise, extracted the NCBI taxonomy, converted iNaturalist observations to RDF, merged observations and taxonomy to create the final ontology, set up krr.triply.cc triplestore, created a documentation with pyLODE, wrote Methodology subparts of the present writing.
- **Sunny Soni** Worked on iNaturalist API and extracted observations data from iNaturalist and GBIF, helped in multiple CSV operations for data preparation, Built web-app with Alex with portion of spread and SPARQL interface, derived optimised way to use haversine distance calculation for large arrays (included in observations_dist.py), worked on results section of the report with Yuyu.
- **Yuyu Bai** Worked on exploring different methods to extract the ontology from the Catalogue of Life (COL), NCBI. Downloaded taxonomy data from the COL and convert it into RDF (sadly we didn't use that in finally ontology). Came up with some interesting queries. Wrote result part and conclusion part and finally made a Docker file.

A.2 Code & Data

All files can be found in the project's Github repository