USING KNOWLEDGE GRAPHS TO CONNECT WILDLIFE INFORMATION SOURCES AND INVESTIGATE IMPACT CAUSED BY WAR.

Maarten van den Ende
2529117
maarten.w.j@gmail.com

Thomas van Zwol
2655625
t.j.vanzwol@gmail.com

June 9, 2019

ABSTRACT

In this research we create a tool to connect the IUCN Redlist data with Wikidata and the WWF Biome information. This is the first and most complete and versatile tool to obtain information about wildlife from different sources. This tool should be able to be used by conservation agencies to discover where and what are the most effective actions to reduce biodiversity reduction.

We then use this as a case study by combining information on active war missions of the UN to see what animals are most threatened by these wars and to discover ending which war has most environmental impact.

1 Introduction

There are many environmental and wildlife protective agencies around the world. The main global ones are World Wildlife Foundation and the International Union for Conservation of Nature. Next to this, many local agencies try to have a positive impact on conserving biodiversity in more specific areas. As an example a quick google search shows over four different organizations specifically focused on preservation of chimpansees, like Chimps Inc., Chimp Haven Inc., Mona Foundation and Save the Chimps. This in addition to more general foundations focused on primates like Primarily Primates Inc. and Ape Cognition and Conservation Project. And then lastly there are the global wildlife organizations like WWF and IUCN mentioned above.

All these organizations work together as much as possible to work as efficiently as possibly on their good cause. IUCN and WWF for example worked together in creating the Wildlife Trade Monitoring Network TRAFFIC [1]. However, they have obtained different data and handle and store their data in different ways. Especially smaller local organizations do not have the ability to easily access the global knowledge of other organizations.

This research has the goal of increasing efficiency of handling data for all agencies, and creating an easier, better accessible and broader knowledge base of wildlife and nature data.

This is done by combining data from WWF and IUCN as well as WikiData into a Knowledge Graph and creating a tool that makes obtaining information from these sources straightforward and easy. This technology is made to be easily extendable with more data sources using existing and specially created ontology with the goal of giving smaller conservation initiatives the possibility to add their specialized knowledge.

As a case study to show the use of this tool we will add the UN Active Warzones data, which uses many features of the tool like the interactive map and flexible queries. We then showcase the capabilities of the tool by answering the question on which war imposes most threats on biodiversity and thus which war UN should focus on when taking nature conservation in account.

2 Method

The methodology of this research in broad terms exists of obtaining data from different sources. This data comes in different formats, with some already being in triple format and other data still had to be converted. To do this we create a custom ontology next to the already existing RDFS, OWL and SKOS. This imposed the biggest challenge in this research. Finally an interactive web app to make accessing and using the data easily obtainable is created.

2.1 Obtaining data

2.1.1 IUCN Redlist Scrape

The main data source used is the IUCN Redlist [2]. IUCN provides an API in which one can query for taxon name, country, region and more. However, the API is heavily focused on species and taxons and as our goal is to create a versatile tool we want to be able to select on much more like type of threaT and conservation measures. To be able to do this we needed to have all data and connect all features using a Knowledge Graph. As a result a scraper was written to obtain all IUCN data of all mammals. Later this could be extended to include all species.

The scraper uses the API to obtain all information per species by going through the list of mammals as provided by the IUCN API. It then takes the information given and puts it into triple format using existing ontology as well as a custom written one which will be discussed in a following section.

As the API is queried per species, scraping just the mammals takes around 10 hours. The obtained dataset in the end exists of the complete data with classification schemes defined by the IUCN Classification Scheme [3].

2.1.2 WWF Data

In this research two datasets from the World Wildlife Foundation are used. They are the Terrestrial Ecoregions Dataset [4] and the Wildfinder Dataset [5]. The Ecoregions dataset, which consists of shapefiles of all ecoregions in the world, is shown in our webapp as a map. This is done by converting it to a GeoJson and shown using Dash and React. The Wildfinder dataset is used to connect the taxon of species with the ecoregions they occur in.

The Wildfinder dataset is a Microsoft Access Database file, consisting of multiple linked data tables. For example one table gives the name of ecoregions for each ecoregion code, while another gives the WWF Species ID with the ecoregions this species occurs in. Then another table connects the WWF ID with the taxon of the animal. Using a

Python script these tables are linked to each other to finally be able to connect the taxon directly with the Ecoregion name they live in.

2.1.3 WikiData

The WikiData Knowledge Base is used to provide the tool with additional general information of the animals. Once the user enters a query the tool requests a SPARQL Query to WikiData using the taxon of the animal. The images and abstract in our web app come from WikiData and prove that the linking is successful. Other data about the species can be incorporated in the web app as they are deemed necessary.

2.1.4 UN Active Missions

The United Nation has a CSV dataset of active missions [6]. It is a small dataset consisting of less than twenty active missions. It gives the name, the location of the UN Headquarters, which can be a village, city or just a general area. It also gives the country code it is located in in multiple formats. The main challenge in this dataset is connecting the locality of the war, as the country codes are unusable, both because of some wars lacking them as well as the lack of data on connecting them with the countries name. This gave us two options: using the GeoNames Database to automatically obtain the countries name or doing it manually. While the automated version would have more scalability we decided that for this few instances the manual solution of taking the longitude and latitude was quicker.

2.2 Ontology and mapping

The IUCN Classification system [3] provides a well defined set of classes and subclasses which could be taken over in the Knowledge Graph. For example, the main class 'Threats' consists of multiple different subclasses like agricultural threats, energy production and pollution. Then these subclasses are more specific in their respective subclasses. For pollution these can be urban waste water or industrial and military effluents. The second one is then classified finally in Oil spills, Seepage from mining and Unknown. The specific threat to a species is an instance of these classes, since the specific threat also has other information that is specific to the species. The same goes for the habitats and measures.

A WWF ontology has been created using the class Species and hasEcoregion and hasGenusSpeciesName have been created to incorporate the ecoregions species live in. The instance hasGenusSpeciesName is then connected using owl:sameAs with the scientific name of the IUCN species.

2.3 Triplestore

We use GraphDB as our triple store. We chose this triple store because it was recommended to us and has an easy to access SPARQL endpoint to query. Furthermore it has a built in reasoner that allowed us to easily link the WWF data with the IUCN data through owl:sameAs links.

2.4 Interactive WebApp

The WebApp consists of the user interface, with easy search queries for users as well as clickable links and clickable biome and warzone map. It is built in React (Javascript) and uses http requests to connect to both the GraphDB database that acts as our triple store and the WikiData SPARQL endpoint.

2.4.1 Linking with database

Connecting both SPARQL endpoints proved a challenge, due to the fact that federated queries to GraphDB didn't work as expected and because sending federated queries to the WikiData endpoint wouldn't make sense since the GraphDB endpoint was on a local machine.

The data from WikiData that is incorporated at this point in time is a proof of concept that linking between the IUCN dataset and WikiData is possible. Other types of data can be incorporated relatively easily in the web app.

2.4.2 User interface

The user interface consists of three main components: the search and select on the left, the information panel in the top right and in the bottom right the interactive ecosystem map. In the search bar one can enter a search term and select what the search term refers to: species name, habitat, threat, ecoregion or measure. Then the possible matches show up as cards with the name of the animal as well as an image. After selecting one, the available data is shown top right, with

clickable links initiating a search for that specific item wherever possible. This makes it so that for example you can explore other animals in the same habitat without querying the habitat manually, but instead by clicking the habitat link.

Lastly there is a map consisting of all terrestrial ecosystems of the world. Hovering shows the name of the ecosystem the cursor is at. Clicking it then gives a query to the tool giving all animals whose habitat is located in that ecosystem. This way it is easy to find and explore different ecosystems and their inhabitants. Next there is a button to change the map to show active war missions on top of the ecosystems. They are shown as red dots that when hovered over provide the name of the war.

3 Results

In this project we have successfully created an interactive tool making it easy for nature and wildlife conservation organizations to access data from WWF and IUCN. The tool works quickly, is more user friendly than the tools given by WWF and IUCN themselves while combining data of both and also incorporating other linked data from WikiData. A system has been put in place to add data relatively easy, which in the future can be used my other organizations to add their specific knowledge and further complete this overview of species and their conservation status.

We have added a simple dataset of Active War Missions of the United Nations with the goal of finding which war is most detrimental to bioconservation. Using this dataset in combination with the interactive tool we deemed that very it is easily possible to find which animals are threatened by each war. However, unfortunately the WWF data lacks ecoregions for mammals. So while the system is in place, unfortunately it could not give a result using only the scraped data for mammals. So our second goal of finding which wars are most detrimental to biodiversity could not be answered.

Results will be updated once the connection discussed in the Discussion is implemented.

4 Discussion

The main focus point for improving our tool would be make finding the effects of wars on nature possible. This can be done to either scrape the IUCN and add more than just mammals, as other animals do have a defined ecoregions in the WWF dataset. Another way of doing this is by finding another source of information that allows linking to the wars. We believe we have identified such a source and are in the process of linking it to the web app: iNaturalist.org. This website is dedicated to identifying plants and animals while at the same time creating data that can be used by researchers. It has an API that allows us to look for the regions where a species has been observed. By linking this to the locations of the peacekeeping missions, it will become possible to identify the animals threatened by these wars.

A limitation to extending our dataset of animals from IUCN, is scraping time. As the API can only be queried per species, the scraper took about 10 hours to run just for mammals. While this is still very viable, it needs to be taken into account when extending for more species, as scraping time might be non-viably long.

The tool as it is now is mainly a proof of concept, with promising abilities. Its use can be increased significantly with added information and utilities.

The map is already a special addition to the tool. More spatially distributed information to be added to the map, like logging work, wildfires or urbanization. The interactivity of the map could be increased as well, making the wars or other data clickable to show more in-depth information.

The WikiData connection is used now to provide additional information. However, this information is readily available and now mainly adds convenience. Connecting the tool with other Knowledge Graphs is a promising proof of concept that could make it even more simple to add additional information owned by local conservation organizations, if they would store their information in a Knowledge Graph.

One nice addition is that the data that is currently being queried from WikiData and that will be queried from iNaturalist.org, can be saved in our triple store as the requests are made. This means that our tool overtime becomes less dependent on the other data sources and our data becomes enriched further.

5 Conclusion

This has been a reasonably successful research project, both in proving a concept and idea as well as giving some useful information that could not be obtained otherwise. We have created an intuitive tool that combines information from multiple sources and with this it could be one of the biggest known datasets on wildlife and its conservation. We have

unfortunately not been able to prove yet that linking these datasets makes it possible to extract findings that otherwise would not be discovered, as lacking WWF data made it impossible to find which mammals were threatened by active wars. Making this possible is our main focus for a future version.

If the tool is extended and used by multiple organizations it could prove to be very useful in aiding nature conservation by making data easier accessible and obtainable and thus streamlining the information process.

References

[1] TRAFFIC. www.iucn.org/news/secretariat/201710/iucn-welcomes-wwf-international-iucn-conservation-centre. Accessed: 2019-04-26.

[2] IUCN Red List API. http://apiv3.iucnredlist.org/. Accessed: 2019-04-26.

[3] IUCN Classification. https://www.iucnredlist.org/resources/classification-schemes. Accessed: 2019-06-03.

[4] WWF Terrestrial Ecoregions. https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world. Accessed: 2019-06-03.

[5] WWF WildFinder. https://www.worldwildlife.org/publications/wildfinder-database. Accessed: 2019-04-26.

[6] United Nations Peacekeeping Missions. https://peacekeeping.un.org/en/peacekeeping-master-open-datasets. Accessed: 2019-06-03.